Householder Symposium XIX
June 8-13, Spa Belgium

Contents

Householder Symposium XIX on Numerical Linear Algebra, June 8-13, 2014 Spa, Belgium

Householder Symposium XIX on Numerical Linear Algebra is the nineteenth in a series, previously called the Gatlinburg Symposia. The series of conferences is named after its founder Alston S. Householder, one of the pioneers in numerical linear algebra.[1] The 2014 symposium is organized in Belgium by the catholic university of Louvain-la-Neuve in collaboration with the Katholieke Universiteit Leuven and the Université de Namur, and sponsored by Mathworks, NAG, FNRS, NSF, ILAS, SIAM and the IAP network DYSCO.

The Householder Symposia originated in a series of meetings organized by Alston Householder, Director of the Mathematics Division of Oak Ridge National Laboratory and Ford Professor at the University of Tennessee. These international meetings were devoted to matrix computations and linear algebra and were held in Gatlinburg, Tennessee. They had a profound influence on the subject. The last "Gatlinburg" conference held at Gatlinburg was in 1969 on the occasion of Householder's retirement. At the time, it was decided to continue the meetings but vary the place. Since then meetings have been held at three-year intervals in a variety of venues and the series has been renamed in honor of Alston Householder. Table 1 contains a complete list of the previous symposia. It has become a tradition to select locations which are relatively remote from possible "distracting activities", and the 2014 selection of Spa makes no exception to this. Even though Spa used to be called the "café of Europe" in the 18th century and was then very touristic, it has nowadays become much "quieter" but it remains well-known for its thermal springs and clean environment.

The meetings, which last for five days, are by invitation only. They are intensive, with plenary talks in the day and special sessions in the evenings. To encourage people to talk about work in progress, no proceedings are published, although extended abstracts are available via the web-page of the symposium. The response of the participants to the meetings has generally been very enthusiastic.

The conferences are run in tandem by a permanent organizing committee and a local arrangements committee. Although attendance is restricted, anyone - including students - can apply. Selection is made by the organizing committee, generally by ballot.

At Householder XIII in Pontresina, Switzerland, F.L. (Fritz) Bauer gave an after-banquet talk, remembering the Symposium's namesake, and the early history of the meetings. Bauer's notes for the talk (posted in NA Digest, July 18, 1996, vol. 96, no. 27) can be found at `http://www3.math.tu-berlin.de/householder_2008/Cleve.html`.

The meeting is also the occasion for the award of the Householder prize for the best thesis in numerical linear algebra. The term numerical linear algebra is intended to describe those parts of mathematical research that have both linear algebraic aspects and numerical content or implications. Thus, for example, a dissertation concerned with the numerical solution of differential equations or the numerical solution of an optimization problem would be eligible if linear algebra is central to the research contribution. This prize is entirely (and well) supported by contributions solicited at the Symposium banquet. The Householder Award, given every three years, was established at the 1969 Gatlinburg Symposium to recognize the outstanding contributions of Alston S. Householder, 1904-1993, to numerical analysis and linear algebra. Nominations are assessed by an international committee. Table 2 contains a complete list of the previous winners.

[1] Prof. G.W. Stewart, University of Maryland, has an article on Alston S. Householder in SIAM News, Vol. 26, October 1993.

Number	**Year**	**Place**	**Organizers**
I	1961	Gatlinburg, U.S.A.	A.S. Householder
II	1963	Gatlinburg, U.S.A.	A.S. Householder, F.W.J. Olver
III	1964	Gatlinburg, U.S.A.	A.S. Householder
IV	1969	Gatlinburg, U.S.A.	A.S. Householder
V	1972	Los Alamos, U.S.A.	R.S. Varga
VI	1974	Hopfen am See, BRD	F.L. Bauer
VII	1977	Asilomar, U.S.A.	G.H. Golub
VIII	1981	Oxford, ENGLAND	L. Fox, J.H. Wilkinson
IX	1984	Waterloo, CANADA	J.A. George
X	1987	Fairfield Glade, U.S.A.	R.C. Ward, G.W. Stewart
XI	1990	Tylosand, SWEDEN	A. Björck
XII	1993	Lake Arrowhead, U.S.A.	T.F. Chan, G.H. Golub
XIII	1996	Pontresina, SWITZERLAND	W. Gander, M.H. Gutknecht, D.P. O'Leary
XIV	1999	Whistler, B.C., CANADA	J.M. Varah, G.W. Stewart
XV	2002	Peebles, SCOTLAND	P. Knight, A. Ramage, A. Wathen, N.J. Higham
XVI	2005	Seven Springs, U.S.A.	J. Barlow, D. Szyld, H. Zha, C. Van Loan
XVII	2008	Zeuthen, GERMANY	J. Liesen, V. Mehrmann, R. Nabben, A. Bunse-Gerstner
XVIII	2011	Tahoe City, U.S.A	Esmond G. Ng, M. Overton

Table 1: List of previous Householder Symposia.

Year	Names
1971	François. Robert (Grenoble)
1974	Ole Hald (New York University)
1977	Daniel D. Warner (University of California, San Diego)
1981	Eduardo Marques de Sa' (Coimbra); Paul Van Dooren (K. U. Leuven) - shared
1984	Ralph Byers (Cornell University); James M. Demmel (University of California, Berkeley) - shared
1987	Nicholas J. Higham (University of Manchester)
1990	Alan Edelman (Massachusetts Institute of Technology) ; Maria Beth Ong (University of Washington) - shared
1993	Hong-Guo Xu (Fudan University) ; Barry Smith (New York University) - shared
1996	Ming Gu (Yale University)
1999	Jörg Liesen (Bielefeld)
2002	Jing-Rebecca Li (Massachusetts Institute of Technology)
2005	Jasper van den Eshof (Utrecht University)
2008	David Bindel (University of California, Berkeley)
2011	Bart Vandereycken (KU Leuven); Paul Willems (Bergische Universität Wuppertal) - shared

Table 2: Previous Householder Award Winners.

Householder Committee

Jim Demmel, University of California at Berkeley, USA
Alan Edelman, Massachusetts Institute of Technology, USA
Heike Fassbender, Technical University of Braunschweig, Germany
Ilse Ipsen (chair), North Carolina State University, USA
Volker Mehrmann, Technical University of Berlin
Jim Nagy, Emory University, USA
Yousef Saad, University of Minnesota, USA
Valeria Simoncini, University of Bologna, Italy
Zdenek Strakos, Charles University in Prague, Czech Republic
Andy Wathen, University of Oxford, UK

Local Organizing Committee

Pierre-Antoine Absil, Université catholique de Louvain, Belgium
Annick Sartenaer, Université de Namur, Belgium
Marc Van Barel, Katholieke Universiteit Leuven, Belgium
Sabine Van Huffel, Katholieke Universiteit Leuven, Belgium
Karl Meerbergen, Katholieke Universiteit Leuven, Belgium
Paul Van Dooren (chair), Université catholique de Louvain, Belgium

Householder Prize Committee

Michele Benzi, Emory University, USA
Inderjit Dhillon, University of Texas Austin, USA
Howard Elman, University of Maryland, USA
Volker Mehrmann (chair), Technical University of Berlin, Germany
Françoise Tisseur, University of Manchester, UK
Stephen Vavasis, University of Waterloo, Canada

Acknowledgments

We are grateful to the following organizations for their generous financial support and sponsorship of Householder Symposium XIX.

International Linear Algebra Society (ILAS)

IAP Network DYSCO (BELSPO)

MathWorks

Numerical Algorithms Group (NAG)

Society for Industrial and Applied Mathematics (SIAM), and SIAM Activity Group on Linear Algebra (SIAG/LA)

Fonds de la Recherche Scientifique, Belgium

U.S. National Science Foundation

Abstracts

(*Abstracts are arranged in alphabetical order of presenters.*)

Algorithms for the Nearest Correlation Matrix Problem with Factor Structure

Charlotte Dorcimont and P.-A. Absil

Abstract

We revisit the problem of computing a nearest correlation matrix with factor structure, addressed by Borsdorf *et al.* in his communication [BHR11] at the 2011 Householder Symposium; see also the related paper [BHR10] and the PhD thesis [Bor12].

The k-factor problem is formulated as follows, where $\mathrm{Diag}(B)$ stands for B with its off-diagonal elements set to zero, $\mathrm{diag}(B)$ is the column vector formed from the diagonal of B, and $\mathbf{1}_n$ denotes the n-dimensional column vector of all ones: Given $A \in \mathbb{R}^{n\times n}$ with unit diagonal and a positive integer $k < n$,

$$\text{minimize } f(X) := \|A - (I + XX^T - \mathrm{Diag}(XX^T))\|_F^2 \tag{1a}$$

$$\text{subject to } X \in \Omega := \{X \in \mathbb{R}^{n\times k} : \mathrm{diag}(XX^T) \leq \mathbf{1}_n\}. \tag{1b}$$

This problem is motivated by the following factor model, which appears in various financial contexts (see [BHR10, §1]):

$$\xi = X\eta + F\varepsilon,$$

where $X \in \mathbb{R}^{n\times k}$, $F \in \mathbb{R}^{n\times n}$ is diagonal, and $\eta \in \mathbb{R}^k$ and $\varepsilon \in \mathbb{R}^n$ are vectors of centered unit-variance independent random variables. The covariance of ξ is

$$\mathrm{cov}(\xi) = \mathbb{E}(\xi\xi^T) = XX^T + F^2.$$

Imposing that $XX^T + F^2$ is a correlation matrix (i.e., a symmetric positive-semidefinite matrix with unit diagonal) yields the constraint (1b), then seeking $XX^T + F^2$ as close as possible to a given A, in the least squares sense, yields the objective function (1a).

Various numerical methods for solving (1) were considered in [BHR10]. Among these methods, the nonmonotone spectral projected gradient method (SPGM) of Birgin, Martínez, and Raydan [BMR00] emerged as the preferred method in terms of efficiency combined with reliability [BHR10, §6]. The simplicity of the projection onto the feasible set Ω partly explains the high efficiency of SPGM on (1). We have also observed in our experiments that the nonmonotone nature of the method plays a crucial role.

In recent work, reported in [Dor13], we have contributed several items to the panoply of methods available to tackle the k-factor problem (1).

One of the new methods reformulates the iterative method of Anderson, Sidenus, and Basu [BHR10, (5.3)] as the zero-finding problem

$$\mathrm{diag}(F(y)F(y)^T) - y = 0$$

$$\text{where } F(y) := \operatorname*{arg\,min}_{X\in\mathbb{R}^{n\times k}} \|A - \mathrm{Diag}(A) - XX^T + \mathrm{Diag}(y)\|_F^2,$$

which we address with a quasi-Newton method. Observe that the minimization reduces to a truncated singular value decomposition. Note that this method, like Anderson's method, does not enforce the constraint (1b), with the hope, sometimes but not always fulfilled, that the algorithm will converge to a feasible point, i.e., a point that satisfies (1b).

Another method proposed in [Dor13], now with feasible iterates, consists of tackling (1) by optimizing over each row of X in turn while keeping the other rows fixed. Each of these subproblems is seen to be a minimization of a quadratic convex function on a unit ball, i.e., an instance of the well-known trust-region subproblem, for which various methods are available. We address those subproblems using the Moré–Sorensen algorithm [MS83]. A significant speedup can be achieved by exploiting the fact that the Hessian of the quadratic function undergoes a low-rank update between successive subproblems.

This row-wise approach has certain features that can be considered advantageous over SPGM: it is finer grained, there are fewer parameters to tune, and we have observed in preliminary experiments that it tends to outperform SPGM in python/numpy implementations.

Finally, we point out that problem (1) is in general multimodal, and the methods mentioned above cannot guarantee more than convergence to a *local* minimizer. The ability to feed the methods with a good initial point may thus lead to considerable improvements. In our preliminary experiments, we have started the methods with the projection onto Ω of the best rank-k approximation of A, but other techniques, notably along the lines discussed in [BHR10, §5], remain to be investigated.

References

[BHR10] Rüdiger Borsdorf, Nicholas J. Higham, and Marcos Raydan. Computing a nearest correlation matrix with factor structure. *SIAM J. Matrix Anal. Appl.*, 31(5):2603–2622, 2010.

[BHR11] Rüdiger Borsdorf, Nicholas J. Higham, and Marcos Raydan. Computing a nearest correlation matrix with factor structure. In *Householder Symposium XVIII on Numerical Linear Algebra*, pages 32–33, 2011.

[BMR00] E. Birgin, J. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000.

[Bor12] Ruediger Borsdorf. *Structured Matrix Nearness Problems: Theory and Algorithms*. PhD thesis, School of Mathematics, The University of Manchester, 2012. MIMS Eprint 2012.63.

[Dor13] Charlotte Dorcimont. Low-rank approximation of correlation matrices. Master's thesis, Ecole Polytechnique de Louvain, Université catholique de Louvain, Av. G. Lemaître 4, 1348 Louvain-la-Neuve, Belgium, 2013.

[MS83] Jorge J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. Sci. Statist. Comput.*, 4(3):553–572, 1983.

Global Convergence of the Restarted Lanczos Method and Jacobi-Davidson Method for Symmetric Eigenvalue Problems

Kensuke Aishima

Abstract

Suppose one wants to compute one or more extremal eigenvalues and their corresponding eigenvectors of a symmetric matrix A. There exist a number of efficient iterative methods for the task. Among them, the Lanczos method [6] is a classical and powerful technique. The Lanczos method performs the so-called Rayleigh-Ritz procedure on a Krylov subspace. In order to reduce the computational and memory costs, a restart strategy [8, 9] should be employed from the practical point of view. However this makes the convergence analysis less straightforward. In this talk, we derive global convergence theorem of the restarted Lanczos method. Key tools for the proof are certain convergence properties of the Rayleigh-Ritz procedure due to Crouzeix, Philippe and Sadkane [2]. As well as Lanczos, there are a number of efficient algorithms that use the Rayleigh-Ritz procedure. Among them, we proved a certain global convergence property of the restarted Jacobi-Davidson method proposed by Sleijpen and van der Vorst [7].

Existing studies of the global convergence of the restarted Lanczos method can be summarized as follows. In 1951, Karush derived global convergence for the restarted strategy to compute one largest eigenvalue [4], and Knyazev and Skorokhodov gave its convergence proof based on certain properties of the steepest descent method [5]. To the best of our knowledge, the most general result about the global convergence is Sorensen's theorem of the implicitly restarted Lanczos method for computing the largest more than one eigenvalues [8]. However, in Sorensen's theorem there is a technical assumption that the absolute values of the off-diagonal elements of the Lanczos tridiagonal matrix are larger than a positive constant throughout the iterations.

In this study, we proved global convergence without any such assumption. The only assumption is that the initial vector is not orthogonal to the exact eigenvectors. In addition, the convergence theorem is extended to restarted Lanczos for computing both the largest and smallest eigenvalues. As for the restarted block Lanczos method, a certain convergence property for a restarted strategy to compute the largest eigenvalues has been already proven by Crouzeix, Philippe and Sadkane [2]. More specifically, the Ritz values converge to exact eigenvalues, although not necessarily to the largest ones. We extend this result to the restart strategy in [3] for computing both the largest and smallest eigenvalues.

Regarding the restarted Jacobi-Davidson method, the local asymptotic convergence rate has been well studied. However, despite many efforts over a decade, the theory for global convergence remains an open problem. In this talk, we show that the largest Ritz value of the restarted Jacobi-Davidson converges to an exact eigenvalue, although not necessarily to the largest one.

It is worth noting that, there are a number of improved versions of the Lanczos and Jacobi-Davidson combined with other effective restarted strategies. However, in this talk, we investigate global convergence properties of the typical and basic versions. We also note that our results cover a dynamic restarting procedure [2], where the restarting points are dynamically determined.

The material of this talk is based on the recent technical report [1].

References

[1] K. Aishima: Global Convergence of the Restarted Lanczos Method and Jacobi-Davidson Method for Symmetric Eigenvalue Problems, *Mathematical Engineering Technical Reports*, METR 2013-27, the University of Tokyo, 2013. (http://www.keisu.t.u-tokyo.ac.jp/research/techrep/)

[2] M. Crouzeix, B. Philippe, and M. Sadkane: The Davidson method, *SIAM J. Sci. Comput.*, vol. 15 (1994), pp. 62–76.

[3] J. K. Cullum: The simultaneous computation of a few of the algebraically largest and smallest eigenvalues of a large, symmetric, sparse matrix, *BIT*, vol. 18 (1978), pp. 265–275.

[4] W. Karush: An iterative method for finding characteristic vectors of a symmetric matrix, *Pacific J. Math.*, vol. 1 (1951), pp. 233–248.

[5] A. V. Knyazev and A. L. Skorokhodov: On exact estimates of the convergence rate of the steepest ascent method in the symmetric eigenvalue problem, *Linear Algebra Appl.*, vol. 154-156 (1991), pp. 245–257.

[6] C. Lanczos: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. Nat. Bur. Standards*, vol. 45 (1950), pp. 255–282.

[7] G. L. G. Sleijpen and H. A. van der Vorst: A Jacobi-Davidson iteration method for linear eigenvalue problems, *SIAM J. Matrix Anal. Appl.*, vol. 17 (1996), pp. 401–425.

[8] D. C. Sorensen: Implicit application of polynomial filters in a k-step Arnoldi method, *SIAM J. Matrix Anal. Appl.*, vol. 13 (1992), pp. 357–385.

[9] K. Wu and H. Simon: Thick-restarted Lanczos method for large symmetric eigenvalue problems, *SIAM J. Matrix Anal. Appl.*, vol. 22 (2000), pp. 602–616.

An Efficient Estimator of the Condition Number of the Matrix Exponential

Awad H. Al-Mohy

Abstract

A new condition number estimator for the matrix exponential is presented. The estimator *avoids* the explicit computation of the Fréchet derivative of the matrix exponential, which underlies condition number estimation in the existing algorithms, so considerable computational savings are possible since computing Fréchet derivative requires twice the cost of the matrix exponential itself.

Introduction

Condition number of a matrix function measures the sensitivity of the matrix function to perturbations in data. The relative condition number of a matrix function $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ at a point $X \in \mathbb{C}^{n\times n}$ is given by [5, Thm. 3.1]

$$\operatorname{cond}(f, X) = \frac{\|L_f(X)\|\,\|X\|}{\|f(X)\|},$$

where $\|L_f(X)\| = \max_{Z\neq 0} \|L_f(X, Z)\|/\|Z\|$ and L_f is a linear operator on $\mathbb{C}^{n\times n}$ defined as

$$f(X + E) - f(X) - L_f(X, E) = o(\|E\|)$$

and is called Fréchet derivative of the matrix function f at the point X in direction E [5, sec. 3.1]. Here X is fixed but E is variable. Thus Fréchet derivative plays a seminal role in computing the condition number of matrix functions. Thus many algorithms have been developed for several matrix function. Al-Mohy and Higham derive an algorithm for simultaneously computing e^X and estimating its condition number [1] and a method to compute the Fréchet derivative of a general matrix function using complex step approximation [2] and recently Al-Mohy, Higham, and Relton derive an algorithm for computing the matrix logarithm and its Fréchet derivative with condition number estimation [3].

Our focuss here is on the matrix exponential. We obtain a new bound relating the condition number of the matrix exponential to the condition number of the squaring phase of the scaling and squaring method (SSM) [5, Ch. 10] and use it to obtain a new estimator with lower computational cost.

A new condition number estimator

The key result in our paper is the following theorem

Theorem 1. *For any nonnegative integer s and $g(X) = X^{2^s}$ we have*

$$\operatorname{cond}(\exp, X) \le \frac{\|2^{-s}X\|_F e^{\|2^{-s}X\|_2}}{\|e^{2^{-s}X}\|_F} \operatorname{cond}(g, e^{2^{-s}X}). \qquad \square \tag{1}$$

The matrix function g represents the squaring phase of the SSM that exploits the identity $e^X = \left(e^{2^{-s}X}\right)^{2^s}$. Note that $\|L_{\exp}(X)\|$ doesn't appear in the bound. We have seen this bound is reasonably sharp in our numerical experiments, however, we found that the quantity

$$\kappa_g(s, X) = \|2^{-s}X\|_F \operatorname{cond}(g, e^{2^{-s}X}) \tag{2}$$

gives better estimate of $\mathrm{cond}(\exp, X)$. Using a collection of test matrices from the matrix exponential literature, we found that the ratio $\kappa_g(s, X)/\mathrm{cond}(\exp, X)$ lies in the interval $[0.97, 4.8]$ whereas the bound in (1) overestimates $\mathrm{cond}(\exp, X)$ within a factor of 46. In addition, we test this ratio for extreme cases by using the multidirectional search method of Dennis and Torczon [6]. A MATLAB code of that algorithm is available in [4] under the name `mdsmax`. We run the maximizer `mdsmax` to seek the maximal value of the ratio $\kappa_g(s, X)/\mathrm{cond}(\exp, X)$ and its reciprocal. The maximizer evaluates this ratio and its reciprocal about 3.7×10^6 times in total towards extreme points using different starting points, and we obtain

$$0.83 \leq \frac{\kappa_g(s, X)}{\mathrm{cond}(\exp, X)} \leq 16.5,$$

indicating that our estimator is reliable.

The key point why we want to avoid the calculation of $L_{\exp}(X, E)$ is that we gain a considerable computational saving when relying on $L_g(e^{2^{-s}X}, E)$. The algorithm of Al-Mohy and Higham [1, Alg. 7.4] computes e^X and estimates $\mathrm{cond}(\exp, X)$. Assuming the cost of computing e^X is μ, the authors point out that 8 Fréchet derivative evaluations are required to estimate the condition number, and the total cost of the algorithm is about 17μ. Since computing $L_{\exp}(X, E)$ costs 2μ [1, Alg. 4.1] and our estimator does not require $L_{\exp}(X, E)$, saving 16μ is possible. Note that when $s = 0$, $\kappa_g(s, X) = \|X\|_F$ since $\mathrm{cond}(g, e^{2^{-s}X}) = 1$ obtained for no cost, and for $s > 0$, $L_{\exp}(X, E)$ and $L_g(e^{2^{-s}X}, E)$ require $12 + 2s$ and $2s$ matrix multiplications, respectively.

References

[1] A. H. Al-Mohy and N. J. Higham. Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation. *SIAM J. Matrix Anal. Appl.*, 30(4):1639–1657, 2009.

[2] A. H. Al-Mohy and N. J. Higham. The complex step approximation to the Fréchet derivative of a matrix function. *Numer. Algorithms*, 53(1):133–148, 2010.

[3] A. H. Al-Mohy, N. J. Higham, and S. D. Relton. Computing the Fréchet derivative of the matrix logarithm and estimating the condition number. *SIAM J. Sci. Comput.*, 35(4):C394–C410, 2013.

[4] N. J. Higham. The Matrix Computation Toolbox. `http://www.ma.man.ac.uk/~higham/mctoolbox`.

[5] N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[6] V. J. Torczon. On the convergence of the multidirectional search algorithm. *SIAM J. Optim.*, 1(1):123–145, 1991.

Model Reduction of Nonlinear Systems in the Loewner Framework

A.C. Antoulas and A.C. Ioniţă

Abstract

The motivation for model order reduction (MOR) stems from an inevitable fact about modern computing. The need for accurate modeling of physical phenomena often leads to large-scale dynamical systems that require long simulation times and large data storage. For instance, one such example is provided by the discretization of partial differential equations over fine grids, which leads to large-scale systems of ordinary differential equations. In these settings, MOR seeks models of low dimension that accurately capture the input-output behavior of the large-scale system while requiring only a fraction of the large-scale simulation time and storage.

A powerful and versatile approach to MOR is provided by the *Loewner framework* for rational interpolation. This approach was introduced in [3] and a major advance was made in [6]. It has since been successfully applied to two main categories of systems: (a) linear systems with multiple inputs and multiple outputs [6, 5], and (b) linear parametric systems [4, 1, 2]. It is the purpose of this talk to present the most recent extension of the Loewner framework to classes of non-linear systems, namely bilinear systems and quadratic non-linear systems.

The Loewner framework is *data-driven* and starts from empirical data. It employs advanced interpolation techniques, overcoming limitations of standard projection methods The empirical data may be provided by physical experimentation or by direct numerical simulation.

• **A brief overview of the Loewner framework**, as it applies to the problem of (tangential) rational interpolation and rational approximation [3, 6]. Consider data consisting of the *right interpolation data* $\{(\lambda_i, \mathbf{r}_i, \mathbf{w}_i) \mid \lambda_i \in \mathbb{C}, \mathbf{r}_i \in \mathbb{C}^{m\times 1}, \mathbf{w}_i \in \mathbb{C}^{p\times 1},\ i = 1, \ldots, k\}$, and of the *left interpolation data* $\{(\mu_j, \ell_j, \mathbf{v}_j) \mid \mu_j \in \mathbb{C}, \ell_j \in \mathbb{C}^{1\times p}, \mathbf{v}_j \in \mathbb{C}^{1\times m},\ j = 1, \ldots, q\}$. The quantities λ_i, μ_j, are points where the underlying function is evaluated, $\mathbf{r}_i$, ℓ_j are referred to as tangential directions on the right and on the left, while $\mathbf{w}_i$, $\mathbf{v}_j$ are right and left tangential values.

The rational interpolation problem aims at finding a rational $p \times m$ matrix function $\mathbf{H}(s)$, expressed in terms of a (descriptor) realization $[\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}]$, i.e. $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$, such that the *right, left constraints* are satisfied: $\mathbf{H}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i$, $\ell_j\mathbf{H}(\mu_j) = \mathbf{v}_j$. The connection of this problem with model reduction lies in the fact that the rational matrix function $\mathbf{H}(s)$, can be considered as the transfer function of the underlying to-be-reduced linear dynamical system.

The key tool for studying this problem is the $q \times k$ *Loewner matrix*, together with the $q \times k$ *shifted Loewner matrix*, associated with the empirical data: $(\mathbb{L})_{ij} = \frac{\mathbf{v}_i\mathbf{r}_j - \ell_i\mathbf{w}_j}{\mu_i - \lambda_j}$, $(\mathbb{L}_\sigma)_{ij} = \frac{\mu_i\mathbf{v}_i\mathbf{r}_j - \ell_i\mathbf{w}_j\lambda_j}{\mu_i - \lambda_j} \in \mathbb{C}$. We also define the quantities $\mathbf{W} = [\mathbf{w}_1\ \cdots\ \mathbf{w}_k] \in \mathbb{C}^{p\times k}$, and $\mathbf{V} = \begin{bmatrix}\mathbf{v}_1^* & \cdots & \mathbf{v}_q^*\end{bmatrix}^* \in \mathbb{C}^{q\times m}$.

The *solution* to the general tangential interpolation/approximation problem is now as follows.

a. If $k = q$ and $(\mathbb{L}_\sigma, \mathbb{L})$, is a regular pencil, then $\mathbf{E} = -\mathbb{L}$, $\mathbf{A} = -\mathbb{L}_\sigma$, $\mathbf{B} = \mathbf{V}$, $\mathbf{C} = \mathbf{W}$, is a minimal realization of an interpolant of the data. Thus, the associated transfer function $\mathbf{H}(s) = \mathbf{W}(\mathbb{L}_\sigma - s\mathbb{L})^{-1}\mathbf{V}$, satisfies the required interpolation conditions. **b.** In the more common case where there are more data than necessary, $(\mathbb{L}_\sigma, \mathbb{L})$ is a singular pencil. Using the basic fact (see [3]) that the (approximate) rank k of $\mathbb{L}$ is equal to the complexity of the underlying system, rank-revealing SVDs of $\mathbb{L}$ and $\mathbb{L}_\sigma$, yield $\mathbf{Y}, \mathbf{X} \in \mathbb{C}^{N\times k}$. The projection then defined by $\mathbf{X}, \mathbf{Y}$, leads to a realization of degree k, of an (approximate) interpolant of the data: $\mathbf{E} = -\mathbf{Y}^*\mathbb{L}\mathbf{X}$, $\mathbf{A} = -\mathbf{Y}^*\mathbb{L}_\sigma\mathbf{X}$, $\mathbf{B} = \mathbf{Y}^*\mathbf{V}$, $\mathbf{C} = \mathbf{W}\mathbf{X}$, i.e. $\ell_j\mathbf{H}(\mu_j) \approx \mathbf{v}_j$, $\mathbf{H}(\lambda_i)\mathbf{r}_i \approx \mathbf{w}_i$.

• **Further developments**. Recently the Loewner framework was extended to deal with linear parameter-dependent systems [4, 1, 2], and even more recently, it has been extended to certain classes of non-linear systems. *It our purpose to discuss this latter case during the meeting.* We conclude this abstract by summarizing the main aspects of the Loewner framework.

• **Model reduction in the Loewner framework: Summary of features**

1. Given input/output or computed data, we can construct with *no computation* (i.e. no factorizations or matrix solves), a singular high order model in generalized (descriptor) state space form. The key tool is the *Loewner pencil*.

2. The philosophy behind this approach is: *collect data and extract desired information.*

3. In applications the singular pencil must be reduced at some stage. This is a natural way for constructing full and reduced models because it does not *force* inversion of $\mathbf{E}$, and moreover it can deal with many input/output ports.

5. In this framework the *singular values* of $\mathbb{L}, \mathbb{L}_\sigma$, offer a *trade-off between accuracy of fit and complexity of the reduced system.*

6. The approach has been xtended to *parametrized systems*, to *bilinear systems* and to *quadtaric nonlinear systems.*

7. At this early stage, we intend to demonstrate the numerical properties of our approach by means of at least the following: reduction of Burgers' equation, reduction of the FitzHugh-Nagumo equations, and reduction of the miscible flow problem as described in [8]. We also intend to make comparisons with the results to [7].

References

[1] A.C. Ionita and A.C. Antoulas, Parametrized model order reduction from transfer function measurements, in Reduced Order Methods for modeling and computational reduction, Series: Modeling, Computations and Applications, A. Quarteroni, G. Rozza (Eds), Springer (2014).

[2] A.C. Ionita and A.C. Antoulas, Data-driven parametrized model reduction in the Loewner framework, submitted to SIAM J. Scientific Computing, March 2013, revised October 2013.

[3] A.C. Antoulas and B.D.O. Anderson, "On the scalar rational interpolation problem," *IMA J. of Mathematical Control and Information*, Special Issue on Parametrization problems, edited by D. Hinrichsen and J.C. Willems, **3**, pp. 61-88 (1986).

[4] A.C. Antoulas, A.C. Ionita, and S. Lefteriu, On two-variable rational interpolation, Linear Algebra and Its Applications, Volume 436: 2889-2915 (2012).

[5] S. Lefteriu, and A.C. Antoulas, A New Approach to Modeling Multiport Systems From Frequency-Domain Data, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, **29**, 14 - 27, Jan. 2010.

[6] A.J. Mayo and A.C. Antoulas, *A framework for the generalized realization problem*, Linear Algebra and Its Applications, Special Issue in honor of P.A. Fuhrmann, Edited by A.C. Antoulas, U. Helmke, J. Rosenthal, V. Vinnikov, and E. Zerz, vol. **425**: 634-662 (2007).

[7] T. Breiten, Interpolatory methods for model reduction of large-scale dynamical systems, PhD Dissertation, Magdeburg, March 2013.

[8] S. Chaturantabut and D.C. Sorensen, Nonlinear Model Reduction via Discrete Empirical Interpolation, *SIAM J. Sci. Comp.* , **32**,(5), 2737-2764, (2010).

A Spectral Analysis of a Discrete two-domain Steklov-Poincaré Operator

Mario Arioli and Daniel Loghin

Abstract

Various analyses going back to the 1980s considered the discrete Steklov-Poincaré operator arising from a regular two-domain decomposition method for linear systems resulting from finite element or finite difference discretizations of model diffusion or, more generally, scalar elliptic problems [1], [2], [3], [4], [5].

It is now known that for uniform discretizations of size h, the condition number of the discrete Steklov-Poincaré operator is of order $O(h^{-1})$. This result is related to the fact that this operator is spectrally equivalent to the discrete square-root Laplacian operator.

In this work we investigate the discrete Steklov-Poincaré operator arising from a domain decomposition method applied to an anisotropic diffusion operator, which also corresponds to the case of a standard diffusion operator discretized on a uniform mesh. We show that the condition number of the Steklov-Poincaré operator is related to the anisotropy present in the problem (diffusion or mesh anisotropy). In turn, this result establishes a spectral equivalence with a fractional (rather than square-root) power of the discrete Laplacian operator. This result is novel and may pave the way to an alternative definition of fractional Sobolev spaces, but also to improved algorithms in domain decomposition technology.

Let $\Gamma \subset \Omega \subset \mathbb{R}^N$ be a $N-1$-dimensional regular manifold. The Steklov-Poincaré operator is defined from the fractional Sobolev space $H_{00}^{1/2}(\Gamma)$ to its dual $H^{-1/2}(\Gamma)$ in the continuous framework. After a finite-element approximation of the Laplace operator in Γ with zero boundary conditions on $\partial\Gamma$, the discrete Steklov-Poincaré operator is spectrally equivalent to the square-root of a special stiffness matrix. For highly anisotropic elliptic operators the trace of the solution on a manifold Γ is still in $H_{00}^{1/2}(\Gamma)$. However, if an anisotropic operator's family, depending on a real parameter, degenerates then the trace could require a different fractional space.

We will discuss the connection between the linear algebra of the fractional power of structured matrices and the finite-element approximation of highly anisotropic elliptic operators in order to give the asymptotic behaviour of the condition number of the matrix representing the Steklov-Poincaré operator approximation.

Let $a \in \mathbb{R}_+$ and let $\Omega = (-1, 1) \times (-a, a)$. Consider the finite element solution of

$$\begin{cases} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{cases}$$

using a partition of Ω into two equal domains separated by a horizontal boundary

$$\Gamma = \{(x, 0) : -1 \le x \le 1\}$$

and subdivided uniformly into equal triangles with sides $h_x = 2/(n+2), h_y = 2a/(n+2)$ with $n = 2m + 1, m \in N$. Let

$$T_k := \text{tridiag}[-1, 2, -1] \in \mathbb{R}^{k\times k}$$

denote a scaled FEM discretisation of $-\mathrm{d}^2/\mathrm{d}x^2$ on a mesh with k interior points and let $I_k \in \mathbb{R}^{k\times k}$ denote the identity matrix. With this notation, the 2D discrete Laplacian matrix $L \in \mathbb{R}^{n^2\times n^2}$ is given by

$$L = \frac{h_x}{h_y} T_n \otimes I_n + \frac{h_y}{h_x} I_n \otimes T_n = \frac{1}{a} T_n \otimes I_n + a I_n \otimes T_n := L_a.$$

The above expression corresponds also to a uniform discretisation of the anisotropic diffusion problem

$$\begin{cases} -(\partial_{xx} + a\partial_{yy})u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{cases}$$

where Ω is any square.

We will refer to L_a above as the *discrete anisotropic Laplacian*.

Let us consider the Schur complement of L_a for a general value of $a > 0$. Let $L_{m,n}$ denote the Laplacian corresponding to a discretisation with m interior nodes in the x-direction, respectively, n in the y-direction. Dropping the subscript a for now, the Laplacian for the original problem is $L = L_{n,n}$ where

$$L_{n,n} = \frac{1}{a}\begin{pmatrix} aL_{m,n} & & -e_m \otimes I_n \\ & aL_{m,n} & -e_1 \otimes I_n \\ -e_m^T \otimes I_n & -e_1^T \otimes I_n & aT \end{pmatrix},$$

$$\text{with} \qquad L_{m,n} = \frac{1}{a}T_m \otimes I_n + aI_m \otimes T_n \quad \text{and} \quad T = \frac{2}{a}I_n + aT_n.$$

Then the Schur complement is (using the block Toeplitz character of $L_{m,n}$)

$$S = T - \frac{1}{a^2}\sum_{i\in\{1,m\}} (e_i^T \otimes I_n)L_{m,n}^{-1}(e_i \otimes I_n) = T - \frac{2}{a^2}(e_m^T \otimes I_n)L_{m,n}^{-1}(e_m \otimes I_n).$$

By a careful analysis of the spectral properties of S, we have the following asymptotic values for the condition number of S when $a = O(h^\theta)$, $\theta \in \mathbb{R}$

$$\kappa_2(S) \sim \begin{cases} O(h^{-2}), & \theta < -1, \\ O(h^{\theta-1}), & \theta \in [-1, 1], \\ O(1), & \theta > 1. \end{cases}$$

Finally, we will illustrate the implication of the previous result on the choice of the preconditioner for S on relevant applications. By using the properties of specific interpolation spaces, we will choose the convenient fractional power of matrices that make the preconditioned problem independent of h and a.

References

[1] P. E. Bjørstad and O. B. Widlund. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM J. Numer. Anal.*, 23(6):1097 – 1120, 1986.

[2] T. F. Chan. Analysis of preconditioners for domain decomposition. *SIAM J. Numer. Anal.*, 24(2):382 – 390, 1987.

[3] M. Dryja. A capacitance matrix method for Dirichlet problem on polygon region. *Numer. Math.*, 39:51–64, 1982.

[4] M. Dryja. A finite element-capacitance method for elliptic problems on regions partitioned into subregions. *Numer. Math.*, 44(2):153 – 168, 1984.

[5] G. H. Golub and D. Mayers. The use of preconditioning over irregular regions. In *Computing methods in applied sciences and engineering, VI (Versailles, 1983)*, pages 3 – 14. North-Holland, Amsterdam, 1984.

Randomized and Quasi-Randomized Algorithms for Low-Rank Approximation of Gram Matrices

Haim Avron, Michael Mahoney, Vikas Sindhwani and Jiyan Yang

Abstract

We consider the problem of constructing a low-rank approximation to the Gram matrices defined by a set of points and a kernel function. More specifically, let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ where $\mathcal{X} \subset \mathbb{R}^d$, and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. The Gram matrix $\mathbf{K}$ is defined by

$$\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) \ .$$

Since k is a positive definite kernel, $\mathbf{K}$ is symmetric positive semi-definite. In many interesting cases, $\mathbf{K}$ is full-rank with a fast decaying spectrum. We are now interested in quickly building a low-rank approximation

$$\mathbf{K} \approx \mathbf{Z}\mathbf{Z}^T$$

where $\mathbf{Z}$ has $s \ll n$ columns.

Motivation. Gram matrices occur in various methods in computational science and related fields. These computations can typically be accelerated using a low-rank approximation, as it replaces operations on a full-rank square matrix (typically an $O(n^3)$ operation) with operations on the low-rank factors (typically an $O(ns^2)$ operation) – at the price of less accurate results.

We now mention a few cases where Gram matrix occur.

- Kernel methods are an important technique in machine learning and related fields of computer science. These methods, typically operate on the Gram matrix, and can be sped up using a low-rank approximation. A well known example is kernel support vector machines.
- Radial basis function interpolation involves solving linear equations involving the Gram matrix corresponding to the basis functions used.
- Numerical solution of the Fredholm integral equation using the Nystrom method gives rise to a linear system involving the Gram matrix. Low rank approximations have been studied for this problem before [2].

Randomized Approximation. Rahimi and Recht recently proposed an elegant and cheap way to approximate the Gram matrix when k is shift-invariant (that is, can be written as $k(x, y) = \phi(x - y)$ for some positive definite function ϕ) [1]. It is based on the celebrated Bochner's theorem, which states that a continuous shift-invariant(scaled) kernel function $k(\mathbf{x}, \mathbf{z}) \equiv k(\mathbf{x} - \mathbf{z})$ is positive definite if and only if it is the Fourier Transform of a probability density p on $\mathbb{R}^d$, i.e., for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$k(\mathbf{x}, \mathbf{z}) = \int_{\mathbb{R}^s} e^{-i(\mathbf{x}-\mathbf{z})^T \mathbf{w}} p(\mathbf{w}) d\mathbf{w} = \mathbb{E}_{\mathbf{w} \sim p} e^{-i(\mathbf{x}-\mathbf{z})^T \mathbf{w}}$$

The integral representation of the kernel may be approximated as follows:

$$\begin{aligned} k(\mathbf{x},\mathbf{z}) &= \int_{\mathbb{R}^d} e^{-i(\mathbf{x}-\mathbf{z})^T\mathbf{w}} p(\mathbf{w}) d\mathbf{w} \\ &\approx \frac{1}{s}\sum_{j=1}^{s} e^{-i(\mathbf{x}-\mathbf{z})^T\mathbf{w}_s} \\ &= \langle \hat{\Psi}(\mathbf{x}), \hat{\Psi}(\mathbf{z}) \rangle_{\mathbb{C}^s} \ , \end{aligned}$$

through the map,

$$\hat{\Psi}(\mathbf{x}) = \frac{1}{\sqrt{s}} \left[e^{-i\mathbf{x}^T\mathbf{w}_1} \dots e^{-i\mathbf{x}^T\mathbf{w}_s} \right] \in \mathbb{C}^s \ .$$

where $\mathbf{w}_1, \dots, \mathbf{w}_s$ are drawn from p.

Let $\mathbf{Z}$ be defined by setting row i to be equal to $\hat{\Psi}(\mathbf{x}_i)$. It can be shown that $\mathbf{Z}\mathbf{Z}^T$ is an entry-wise approximation to $\mathbf{K}$ [1]. Note that the construction just described is complex; there are also real variants.

Approximations for Semigroup Kernels. While shift-invariant kernels are an important family of kernels, there are other families of kernels that are frequently used. We will present a method to approximate the Gram matrix corresponding to kernels on $\mathbb{R}^d_+$ that are sum-invariant, i.e $k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}+\mathbf{y})$ for some positive definite ϕ. Such kernels occur frequently in computer vision.

Improved Approximations using Quasi-Randomness. For a good approximation Rahimi and Recht's method requires a very large s. The underlying reason is that the method uses a *Monte-Carlo* approximation to the integral. We discuss the use of *Quasi-Monte Carlo* instead: the use of a deterministic quasi-random (low-discrepancy) sequences instead. We will present a theoretical characterization of the approximation error, as well as a method to numerically optimize sequences to produce better low-rank approximation. We will present experimental results using classical low discrepancy sequences (e.g. Halton and Sobol), and using optimized sequences.

References

[1] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems (NIPS)*, 2007.

[2] Lothar Reichel. Fast solution methods for Fredholm integral equations of the second kind. *Numerische Mathematik*, 57(1):719–736, 1990.

Variational Principles and Scalable Solvers for the Linear Response Eigenvalue Problem

Zhaojun Bai, Ren-Cang Li, Dario Rocca and Giulia Galli

Abstract

The linear response eigenvalue problem (LREVP) is of the form

$$\begin{bmatrix} A & B \\ -B & -A \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} \Sigma & \Delta \\ \Delta & \Sigma \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix},$$

where A, B and Σ are symmetric matrices and Δ is a skew-symmetric matrix. Furthermore, $A \pm B$ are positive definite and $\Sigma \pm \Delta$ are nonsingular. It can be shown that the eigenvalues of LREVP are real and appear in pairs $\pm\lambda$. The LREVP is a special eigenvalue problem for so-called doubly structured matrix pencils [1,2].

As an application in quantum mechanics, the LREVP appears as the computational kernel in the linear response perturbation analysis in the time-dependent density functional theory (TDDFT) for excited states calculations [3]. Excited states calculation is a important tool in the study of collective motion of many particle systems, ranging from silicon nanoparticles to the analysis of interstellar clouds. There are immense interests in developing efficient numerical algorithms and simulation techniques for excitation state calculations of molecules and solids for materials design in energy science.

Due to the nature of the linear response theory, the dimension of the LREVP is typically very large. For example, a plane-wave based calculation for the excited energies of fullerence C_{60} in QUANTUM EXPRESSO leads to an LREVP of the order 22 millions [4]. In practice, the first few smallest positive eigenvalues and the corresponding eigenvectors are of particular interests.

Since the linear response theory was proposed for studying the collective motion of many particles in the early 1950s, the development of numerical methods for solving the LREVP has been an active research subject in computational (quantum) physics and chemistry, and in numerical analysis community for over five decades. A 2009 survey study compared Lanczos, Arnoldi, Davidson, and conjugate gradient methods and discussed the limitations of each of these methods for developing an efficient scalable eigensolver [5]. In the study, severe limitations were experienced for Lanczos, Arnoldi and Davidson methods due to the orthogonality constraints, for the CG methods to compute several eigenpairs simultaneously, and for incorporating preconditioning techniques.

In the past three years, we have participated a synergistic study of theory, computation and applications of LREVPs. We uncovered new variational principles to characterize the eigenvalues of interest and Cauchy-like interlacing inequalities between the eigenvalues of the original and reduced LREVPs [6,7]. Although the LREVP is a non-Hermitian eigenvalue problem, these theoretical results mirror the well-known variational principles of the Hermitian eigenvalue problem. With the help of these new theoretical results, we are able to develop the *best* (possible) approximations of few smallest positive eigenvalues via structure-preserving projection, and derive locally optimal steepest descent (SD) and conjugate gradient (CG) algorithms, based on a novel 4D search idea instead of the usual line-search [8,9]. Furthermore, the new locally optimal block 4D SD and CG algorithms allow us to use blocking strategies to perform multiple computation steps of the algorithm for each communication step and to incorporate proper preconditioners for fast convergence. The new algorithms are memory-efficient, and numerical convergence behavior is less stringent on

the (orthogonal-like) normalization constraints of projection subspaces. A successful excited state calculation of the order 22 millions of the LREVP for fullerene C_{60} by our methods was recently presented in [4].

We are working on parallel computation and communication strategies under the communication avoiding algorithmic design paradigm. On modern computer architectures, communication costs in terms of the performance of an algorithm can be much greater than arithmetic computation costs, and the gap is only going to widen in future systems [10]. One of our research issues is how to efficiently computing matrix-vector and matrix-matrix operations by exploiting the matrix (sub)-structures and sparsity of the LREVP. Another issue is how to hide costly global synchronization latency in the 4D search SD and SG algorithms. This is in fact a challenging issue for the scalability of all preconditioned CG and Krylov subspace type eigensolvers.

References

[1] C. Mehl, V. Mehrmann and H. Xu, Canonical forms for doubly structured matrices and pencils. Elec. J. Lin. Alg. 7, pp.112-151, 2000

[2] C. Mehl, V. Mehrmann and H. Xu, On doubly structured matrices and pencils that arise in linear response theory. Lin. Alg. Appl. 380, pp.3-51, 2004.

[3] M. E. Casida. Time-dependent density-functional response theory for molecules. In D.P. Chong, editor, *Recent advances in Density Functional Methods*, pp.155-189, World Scientific, Singapore, 1995.

[4] D. Rocca. Iterative diagonalization of non-Hermitian eigenproblems in time-dependent density functional and many-body perturbation theory, APS Annual Meeting, Session B39, Boston, 2012.

[5] S. Tretiak, C. M. Isborn, A. M. N. Niklasson, and M. Challacombe. Representation independent algorithms for molecular response calculations in time-dependent self-consistent field theories. *J. Chem. Phys.*, 130(5), p.054111, 2009.

[6] Z. Bai and R.-C. Li, Minimization principles of the linear response eigenvalue problem I: Theory, *SIAM J. Matrix Anal. Appl.*, 33(4), pp.1075-1100, 2012.

[7] Z. Bai and R.-C. Li, Minimization principles and computation of the generalize linear response eigenvalue problem, submitted, 2013

[8] D. Rocca, Z. Bai, R.-C. Li and G. Galli, A block variational procedure for the iterative diagonalization of non-Hermitian random-phase approximation matrices, *J. Chem. Phys.*, 136, p.034111, 2012.

[9] Z. Bai and R.-C. Li, Minimization principles of the linear response eigenvalue problem II: Computation, *SIAM J. Matrix Anal. App..*, 34(2), pp.392-416, 2013.

[10] J. Demmel, M. Hoemmen, M. Mohiyuddin, and K. Yelick. Avoiding communication in sparse matrix computations. in *Proceedings of IEEE International Symposium on Parallel and Distributed Processing*, pp.1-12, 2008.

Reconstructing Householder Vectors from Tall-Skinny QR

Grey Ballard, James Demmel, Laura Grigori, Mathias Jacquelin, Hong Diep Nguyen, and Edgar Solomonik

Abstract

Because of the rising costs of communication (*i.e.*, data movement) relative to computation, so-called *communication-avoiding* algorithms—ones that perform roughly the same computation as alternatives but significantly reduce communication—often run with reduced running times on today's architectures. In particular, the standard algorithm for computing the QR decomposition of a tall and skinny matrix (one with many more rows than columns) is often bottlenecked by communication costs. A more recent algorithm known as Tall-Skinny QR (TSQR) is presented in [3] and overcomes this bottleneck by reformulating the computation. In fact, the algorithm is communication optimal, attaining a lower bound for communication costs of QR decomposition (up to a logarithmic factor in the number of processors) [2]. Not only is communication reduced in theory, but the algorithm has been demonstrated to perform better on a variety of architectures, including multicore processors, graphics processing units, and distributed-memory systems.

The standard algorithm for QR decomposition, which is implemented in libraries like LAPACK, ScaLAPACK, and Elemental, is known as Householder-QR. For tall and skinny matrices, the algorithm works column-by-column, computing a Householder vector and applying the corresponding transformation for each column in the matrix. When the matrix is distributed across a parallel machine, this requires one parallel reduction per column. The TSQR algorithm, on the other hand, performs only one reduction during the entire computation. Therefore, TSQR requires asymptotically less inter-processor synchronization than Householder-QR on parallel machines (TSQR also achieves asymptotically higher cache reuse on sequential machines).

Computing the QR decomposition of a tall and skinny matrix is an important kernel in many contexts, including standalone least squares problems, eigenvalue and singular value computations, and Krylov subspace and other iterative methods. In addition, the tall and skinny factorization is a standard building block in the computation of the QR decomposition of general (not necessarily tall and skinny) matrices. In particular, most algorithms work by factoring a tall and skinny panel of the matrix, applying the orthogonal factor to the trailing matrix, and then continuing on to the next panel. Although Householder-QR is bottlenecked by communication in the panel factorization, it can apply the orthogonal factor as an aggregated Householder transformation efficiently, using matrix multiplication.

The Communication-Avoiding QR (CAQR) [3] algorithm uses TSQR to factor each panel of a general matrix. One difficulty faced by CAQR is that TSQR computes an orthogonal factor that is implicitly represented in a different format than that of Householder-QR. While Householder-QR represents the orthogonal factor as a set of Householder vectors (one per column), TSQR computes a tree of smaller sets of Householder vectors (though with the same total number of nonzero parameters). In CAQR, this difference in representation means that the trailing matrix update must be done using the implicit tree representation rather than matrix multiplication as is possible with Householder-QR. From a software engineering perspective, this means writing and tuning more complicated code. Furthermore, from a performance perspective, the update of the trailing matrix within CAQR is less communication efficient than the update within Householder-QR by a logarithmic factor in the number of processors.

Building on a method introduced by Yamamoto [4], we show in [1] that the standard Householder vector representation may be recovered from the implicit TSQR representation for roughly the

same cost as the TSQR itself. The key idea is that the Householder vectors that represent an orthonormal matrix can be computed via LU decomposition (without pivoting) of the orthonormal matrix subtracted from a diagonal sign matrix. That is, given an $m \times b$ orthonormal matrix Q, the Householder-QR algorithm produces a factorization

$$Q = \left(\begin{bmatrix} I \\ 0 \end{bmatrix} - \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} T Y_1^T \right) S$$

where $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ stores the Householder vectors, Y_1 is $b \times b$ and unit lower triangular, T is $b \times b$ and upper triangular (and can be computed from Y), and S is a diagonal sign matrix. This implies that

$$\begin{bmatrix} S \\ 0 \end{bmatrix} - Q = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \left(T Y_1^T S \right),$$

which is the unique LU decomposition. In the talk we will explain how this idea leads to a communication-efficient method for reconstructing the Householder vectors Y from a matrix Q computed from TSQR. We will sketch the proof that our reconstruction method is as numerically stable as Householder-QR (independent of the matrix condition number) and validate this proof with experimental results.

This reconstruction method allows us to get the best of the TSQR algorithm (avoiding synchronization) as well as the best of the Householder-QR algorithm (efficient trailing matrix updates via matrix multiplication). By obtaining Householder vectors from the TSQR representation, we can logically decouple the block size of the trailing matrix updates from the number of columns in each TSQR. This abstraction makes it possible to optimize the panel factorizations and the trailing matrix updates independently. We will present experimental results to demonstrate that our parallel implementation outperforms ScaLAPACK, Elemental, and our own lightly-optimized CAQR implementation on the Hopper Cray XE6 platform at NERSC. While we do not experimentally study sequential performance, we expect our algorithm will also be beneficial in this setting, due to the cache efficiency gained by using TSQR.

References

[1] G. Ballard, J. Demmel, L. Grigori, M. Jacquelin, H. D. Nguyen, and E. Solomonik. Reconstructing Householder vectors from tall-skinny QR. Technical report, EECS Department, University of California, Berkeley, 2013.

[2] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in numerical linear algebra. *SIAM Journal on Matrix Analysis and Applications*, 32(3):866–901, 2011.

[3] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM Journal on Scientific Computing*, 34(1):A206–A239, 2012.

[4] Y. Yamamoto. Aggregation of the compact WY representations generated by the TSQR algorithm, 2012. Conference talk presented at SIAM Applied Linear Algebra.

Block Gram-Schmidt Downdating

Jesse L. Barlow

Abstract

The problem of deleting multiple rows from a Q-R factorization is considered. In that context, suppose $X \in \mathbb{R}^{m\times n}, m > n$, has the factorization

$$X = UR \tag{1}$$

where $U \in \mathbb{R}^{m\times n}$ is left orthogonal and $R \in \mathbb{R}^{n\times n}$ is upper triangular. Suppose X is partitioned

$$X = \begin{matrix} p \\ m-p \end{matrix} \begin{pmatrix} X_0 \\ \bar{X} \end{pmatrix} \tag{2}$$

for some integer p where $m \geq n + p$ and that it is desired to efficiently produce the Q-R decomposition

$$\bar{X} = \bar{U}\bar{R}. \tag{3}$$

Obtaining (3) inexpensively from (1), called the *block downdating* problem, is important in the context of solving recursive least squares problem where observations are added or deleted over time.

The problem is solved by choosing a left orthogonal matrix $B \in \mathbb{R}^{m\times p}$ such that

$$B = \begin{matrix} p \\ m-p \end{matrix} \begin{pmatrix} V \\ 0 \end{pmatrix}$$

where $V \in \mathbb{R}^{p\times p}$ is orthogonal, producing a triple $Q_B \in \mathbb{R}^{m\times p}$, $S_B \in \mathbb{R}^{n\times p}$, and $R_B \in \mathbb{R}^{p\times p}$ such that

$$\begin{aligned} B &= US_B + Q_B R_B & (4)\\ 0 &= U^T Q_B & (5) \end{aligned}$$

where Q_B is left orthogonal and R_B upper triangular. The next step is to find an orthogonal $Z \in \mathbb{R}^{(n+p)\times(n+p)}$ such that

$$Z^T \begin{pmatrix} R_B & 0 \\ S_B & R \end{pmatrix} = \begin{pmatrix} R_V & Y_0 \\ 0 & \bar{R} \end{pmatrix}$$

where $\bar{R}$ remains upper triangular and is the desired upper triangular matrix in (3). The left orthogonal matrix $\bar{U}$ in (3) is recovered from

$$\begin{pmatrix} Q_B & U \end{pmatrix} Z.$$

The matrices V, Q_B, S_B and R_B are produced by a block Gram-Schmidt algorithm. However, difficulties occur when $\begin{pmatrix} B & U \end{pmatrix}$ is nearly rank deficient. In that case, although the condition (4) is straightforward to enforce, the condition (5) is not. Developing an algorithm that keeps both $\|B - US_B - Q_B R_B\|_2$ and $\|U^T Q_B\|_2$ small enough to produce a stable block downdate is the key issue in this talk.

Diffusion Models for Covariance

Christopher Beattie

Abstract

The prediction and estimation of spatially varying random processes is an important component of geologic modeling, weather prediction, and ocean modeling. Both variational data assimilation and statistical interpolation (kriging) methods depend on knowing the covariance structure of quantities of interest. Typically, this covariance structure is unknown, yet in principle, it may be estimated from the same data that is used to predict or estimate the main quantities of interest. In many cases, especially for nonstationary random processes that commonly occur in geophysical and atmospheric sciences and oceanography, the available data are too sparse in either space or time to provide reliable estimates, so additional structural features must be assumed and included to bridge the gap.

A recent approach to this problem begins with the remarkable observation that two classes of covariance kernels commonly used in spatial statistics, the Gauss and the Matérn classes, may be identified with Green's functions of differential operators. This interpretation allows for straightforward modeling of anisotropy and nonhomogeneity through low order parameterization of associated diffusion coefficients. More importantly, this approach introduces the possibility of leveraging the versatile strategies developed for the numerical treatment of PDEs into computationally efficient (implicit) representations of the associated grid-point correlation functions. This is critical for problems of large dimension where explicit representation of grid-point correlation can lead rapidly to unmanageable computational bottlenecks.

I will present some background for this modeling paradigm within the context of oceanographic data assimilation and discuss two open problems that remain: renormalization (estimating the "diagonal" of the kernel) and treatment of boundary conditions. Preconditioned iterative methods developed for solving elliptic boundary problems will be seen to inform useful new approaches for manipulating these correlation models.

The Riccati Eigenproblem

Peter Benner and Ludwig Kohaupt

Abstract

The stabilization of linear dynamical systems is a central task in control theory. One of the most successful approaches is based on the linear-quadratic regulator (LQR) design which finds a stabilizing feedback that at the same time is optimal with respect to a quadratic cost functional. It has proved its robustness and reliability in many real-world applications since its invention by Kalman in 1960, not the least important of the real-world applications certainly was the Apollo 11 mission to the Moon in 1969. Mathematically formulated, one has to solve the problem (here posed in its simplest form, versatile generalizations can be found in the literature):

$$\min_{u\in L_2[0,\infty]} \int_0^\infty x(t)^T Q x(t) + u(t)^T R u(t)\, dt \tag{1}$$

subject to

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0. \tag{2}$$

Here, $x(t) \in \mathbb{R}^n$ is the state of the linear system with given initial value $x_0 \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ is the sought-after control function minimizing the cost functional (1), and $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, Q = Q^T \in \mathbb{R}^{n\times n}, R = R^T \in \mathbb{R}^{m\times m}$. Assuming Q positive semi-definite, R positive definite, and (A, B) stabilizable, it is known that the optimal control function solving (1)–(2) is given by the *Riccati feedback*

$$u(t) = -R^{-1}B^T X_* x(t), \tag{3}$$

where $X_* = X_*^T \in \mathbb{R}^{n\times n}$ is the unique positive semidefinite solution of the *algebraic Riccati equation*

$$0 = Q + A^T X + XA - XBB^T X. \tag{4}$$

The so-obtained optimal feedback is *stabilizing* in the sense that when inserted into (2), it holds $x(t) \to 0$ $(t \to \infty)$ for any initial value $x_0 \in \mathbb{R}^n$ as $A - BB^T X_*$ has all its eigenvalues in the open left half of the complex plane. Therefor, X_* is also called *stabilizing*.

Despite its success in practice, the Riccati feedback approach has drawbacks: it often leads to oscillatory behavior and overshoot in the transient pase, in particular when applied to mechanical systems. Inspired by the successful application of the Lyapunov eigenvalue problem to the analysis of linear systems in [1], we derive the following *Riccati eigenproblem*

$$\mu X = A^T X + XA - XBB^T X. \tag{5}$$

Here, μ is a scalar quantity that we interpret as an eigenvalue, with the corresponding eigenmatrix X. This is a nonlinear eigenvalue problem with nonlinearity in the eigenvector/-matrix.

Assuming a solution (μ, X_*) to (5) with real $\mu < 0$ and $0 \neq X_\# = X_\#^T$ positive semidefinite, the eigenmatrix can now replace X_* in (3). We will see that the so-obtained feedback solution stabilizes the system and often exhibits a better transient behavior than the usual Riccati feedback, avoiding overshoot and oscillations in mechanical systems. We discuss the existence of eigenpairs solving (5) and having the desired properties, and derive a first numerical algorithm to compute such eigenpairs. We will point out open problems as well as possible improvements regarding the numerical algorithm.

References

[1] L. Kohaupt. Solution of the matrix eigenvalue problem $VA + A^*V = \mu V$ with applications to the study of free linear systems. *J. Comp. Appl. Math.* 213(1):142–165, 2008.

Numerical Analysis of Quantum Graphs

Mario Arioli and Michele Benzi

Abstract

A *quantum graph* is a graph where we associate with each edge a differential law that models the interaction between the two corresponding vertices. The use of quantum graphs (as opposed to more elementary graph models, such as simple unweighted or weighted graphs) opens up the possibility of modeling the interactions between agents identified by the graph's vertices in a far more detailed manner than with standard graphs. Quantum graphs are now being widely used in physics, chemistry and engineering (nanotechnology) problems but can also be used, in principle, in the analysis of complex phenomena taking place on large complex networks, including social and biological networks. Such graphs are characterized by highly skewed degree distributions, small diameter, high clustering coefficients and have topological and spectral properties that are quite different from those of the highly regular graphs or lattices arising in physics and chemistry applications.

The main purpose of this talk is to introduce the audience to the numerical analysis of PDEs posed on quantum graphs. We propose numerical methods, based on finite elements and an appropriate ordering of the graph vertices, for the solution of PDEs associated with quantum graphs. We consider simple elliptic, parabolic, hyperbolic and Schrödinger-type equations and we show how these can be approximated to yield large-scale systems of algebraic equations (or, in the time-dependent case, systems of ODEs) which can be solved by preconditioned iterative methods (resp., Krylov-based exponential integrators). In particular, we show that the usual graph Laplacian can be seen as an approximation of a more general Laplace operator acting on a quantum graph.

The long-term goal of the project is to investigate the numerical solution of PDEs posed on quantum graphs with complex topologies. As a simple example, we consider a simple diffusion model for the spreading of information on a complex network of interacting agents.

Can Numerical Linear Algebra make it in Nature?

Paolo Bientinesi, Diego Fabregat and Yurii Aulchenko

Abstract

Among all the scientific journals, the collection "Nature-*" enjoys some of the highest impact factors. In many disciplines, having a publication in such journals is a prestigious and even career-defining event. Although their focus is said to be "on research results of outstanding importance for an interdisciplinary readership", contributions from computer science and mathematics are hardly ever published. This work is meant to illustrate how numerical linear algebra, in its supporting—but critical—role for genome analysis, has a realistic shot at a Nature journal.

Genome-Wide Association Studies (GWAS) are a powerful statistical tool to analyze the relationship between variations in the DNA sequence and complex traits (e.g., diseases). Since the first appearance of GWAS in 2005 and 2006, the amount of published studies increased steadily, well exceeding 4000 articles in 2012. (Several GWAS-related paper appeared in one of the Nature journals.) An effective approach to perform GWAS is based on linear mixed models (LMM), and relies on the processing of terabytes of data while solving billions, or even trillions, of generalized least-squares (GLS) problems: $b := \left(X^T M^{-1} X\right)^{-1} X^T M^{-1} y$.

In actual studies, one has to solve a two-dimensional grid of $m \times t$ GLS problems of size n, where m and t are of the order of millions and hundreds of thousands, respectively. The complexity for the solution of one GLS problem in isolation is $O(n^3)$ floating point operations (flops), adding up to $O(mtn^3)$ flops for the entire study. In one of our experiments, we considered $m = 10^6$, $t = 10^5$, and $n = 10^3$, resulting in roughly 10^{20} flops to be executed. No matter how optimized a GLS solver is, such a GWAS would be kept within hours only by using in its entirety one of the fastest supercomputers in the world; meanwhile, biologists aim at performing GWAS hundreds of times larger.

Thanks to a close collaboration with computational biologists, we were able to uncover the correlations among successive GLS problems, and the structure underlying each of them. In detail, the GWAS can be expressed as

$$\begin{aligned} & b_{ij} := \left(X_i^T M_j^{-1} X_i\right)^{-1} X_i^T M_j^{-1} y_j, \\ & \text{with } X_i = [X_L | X_{Ri}], \text{ and } M_j = \sigma_j(\Phi + h_j I), \\ & \text{where } i = 1, \ldots, m, \text{ and } j = 1, \ldots, t. \end{aligned}$$

Moreover, t is either 1 or $\approx 10^5$. By tackling the GWAS as a whole, i.e. all the $m \times t$ problems at once, by tracking the dependencies between adjacent GLS', and by exploiting their structure, we reduced the computational complexity down to $O(mtn)$. This was the first, crucial step towards making large scale GWAS feasible.

Beyond the computation complexity, GWAS is also challenging because of the size of the datasets. Any large analysis entails the processing of terabytes of data. Specifically, the aforementioned analysis (of modest size) involves reading vectors and matrices for 10 GBs, and writing a tensor of 3.2 TBs. In general, the storage requirements largely exceed even the combined main memory of current typical clusters, and demand an out-of-core mechanism to efficiently handle datasets residing on disk.

In this scenario, data must be partitioned in *slabs*, and it becomes critical to determine the most appropriate traversal, as well as the size and shape of each slab. The problem is far from trivial, because these factors affect the amount of transfers and computation performed, as well as the necessary extra storage. By modeling all these factors, we found the best traversal direction, and determined the shape and size of the slabs to achieve a complete overlap of transfers with computations. As a result, irrespective of the data size, the efficiency of our in-core solver is sustained.

In summary, by designing parallel algorithms that take full advantage of the properties inherent to the generalized-least squares problems as they appear within GWAS, we allowed biologists to perform analysis of arbitrary size, while reducing the computation time by a factor of 1000.
This work is currently under review for publication in Nature Methods.

Music of the Microspheres: from Eigenvalues Perturbations to Gyroscopes

David Bindel and Erdal Yilmaz

Abstract

In 1890, G. H. Bryan demonstrated that when a ringing wine glass rotates, the shape of the vibration pattern precesses. Today, this effect is the basis for a family of high-precision solid-wave gyroscopes widely used in spacecraft. These devices effectively sense rotation by how it perturbs a degenerate mode pair. In recent miniaturized solid-wave gyroscopes, geometry distortions due to imperfect fabrication also perturb the dynamics, and this limits sensing accuracy. In this talk, we describe how geometric imperfections affect the dynamics of solid wave gyroscopes from the perspective of group theory and eigenvalue perturbation theory.

Free vibrations of an ideal solid-wave gyroscope are described by a finite element model

$$M\ddot{u} + 2\Omega B\dot{u} + Ku = 0,$$

where M and K are symmetric and positive definite matrices representing mass and stiffness, and B is skew-symmetric matrix corresponding to the Coriolis effect. By using shape functions that are trigonometric functions in θ, we can put M, B, and K in block diagonal form, where each block corresponds to a different azimuthal wave number. Under usual operating conditions, the Coriolis term is a small perturbation, and the vibrations of interest can be written as

$$u \approx u^1 q_1(t) + u^2 q_2(t),$$

where u^1 and u^2 are a pair of degenerate mode shapes for the structure in an inertial frame in which Coriolis forces are absent. This leads to the two-degree of freedom model

$$\ddot{q} + 2\beta\Omega J\dot{q} + \omega_0^2 q = 0.$$

In the two-dimensional configuration space, solutions to this equation trace the shape of a precessing ellipse, where the rate of rotation of the ellipse is $\beta\Omega$. The imperfections that hurt performance are those that break the degeneracy of the double eigenvalue ω_0, and hence qualitatively change this configuration space picture.

We write perturbations of an ideal geometry in cylindrical coordinates as $(r, z) \mapsto (r, z) + \epsilon f(r, z, \theta)$, and we reason about how the frequencies change with ϵ based on a Fourier series for f. Many of the geometric distortions due to microfabrication issues are highly structured, and so only involve a small number of Fourier modes. Mode frequencies at azimuthal wave number m can change at first order in ϵ only if the Fourier series for f has a nonzero coefficient at azimuthal number $p = 2m$. Past first order, to compute the effect of a p-fold rotationally symmetric perturbation on a mode with azimuthal number m in the unperturbed geometry, we expand the perturbed motion in a Fourier series in θ; selection rules guarantee that the expansion terms at wave number n may only be nonzero if p divides $m \pm n$. Using this approach, we are able to predict qualitative trends, such as the relative insensitivity of modes with azimuthal number 3 to typical microfabrication issues. We also describe how this structure leads to fast algorithms for computing the perturbed frequencies and mode shapes for specific geometric deformations.

References

[1] E. Yilmaz and D. Bindel, *Effects of Imperfections on Solid-Wave Gyroscope Dynamics*, Proceedings of IEEE Sensors, Baltimore, MD, Nov 2013.

Block-smoothing in Multigrid Methods for Circulant and Toeplitz Matrices

Matthias Bolten

Abstract

Circulant matrices and Toeplitz matrices arise in a variety of applications including signal processing and partial differential equations. The solution of systems where the system matrix is a (multilevel) circulant matrix can be obtained with the help of the FFT and thus in $\mathcal{O}(N \log N)$ operations, where N is the number of unknowns. The solution of systems where the system matrix is multilevel Toeplitz is not as easily possible. While iterative methods like the CG method provide a viable alternative, in many applications the matrices are illconditioned, resulting in a rising number of iterations when the system size is increased.

For this reason the development of multigrid methods as efficient iterative methods that do not show this behavior for an important subclass of these matrices have been developed, e.g., in [4] for Toeplitz matrices and in [5] for multilevel circulant matrices. These developements have been continued and it was possible to show optimality of these methods at least in the case of certain matrices from matrix algebras [1]. Our recent work focussed on the choice of grid transfer operators [2, 3]. Most of the analysis focusses on Richardson iterations as smoother. This is sufficient from a theoretical viewpoint, nevertheless a different choice of the smoother can significantly speed up the resulting methods.

Possible choices for smoothers include Gauss-Seidel or SOR, optionally with different colorings, further line- or plane-smoothers can be used, as well as incomplete factorizations. These methods, including Richardson, have in common that the amount of work that has to be carried out is relatively small compared to the number of unknowns and thus compared to the memory transfers that are necessary. The arithmetical intensity can be increased in different ways, one option being the use of block smoothers, where small d-dimensional blocks are inverted instead of the relaxation of single unknowns. The increased arithmetic intensity results in a better utilization of modern computer architectures, like accelerators or parallel computers. As a side effect the efficiency of the smoother is improved. As a result, while the number of FLOPS might go up, the total time to solution can still be better than in the case of block smoothers.

We will present recent results on the usage of block smoothers in multigrid methods for circulant and Toeplitz matrices. For the theoretical analysis proper decompositions of the matrices are used that allow for a detailed analysis of the resulting method. The analysis tools fit in the established framework that is used to analyze multigrid methods for Toeplitz and circulant matrices, allowing for a rigorous analysis of the resulting methods.

Practical results obtained on recent computer architectures demonstrate the efficacy of the chosen approach to reduce the time to solution.

References

[1] A. Aricò and M. Donatelli. A V-cycle multigrid for multilevel matrix algebras: proof of optimality. *Numer. Math.*, 105:511–547, 2007.

[2] M. Bolten, M. Donatelli, and T. Huckle. Analysis of smoothed aggregation multigrid methods based on Toeplitz matrices. *Preprint BUW-IMACM 13/10*, 2013.

[3] M. Bolten, M. Donatelli, T. Huckle, and C. Kravvaritis. Generalized grid transfer operators for multigrid methods applied on Toeplitz matrices. *Preprint BUW-IMACM 13/11*, 2013.

[4] G. Fiorentino and S. Serra. Multigrid methods for Toeplitz matrices. *Calcolo*, 28:238–305, 1991.

[5] S. Serra-Capizzano and C. Tablino-Possio. Multigrid methods for multilevel circulant matrices. *SIAM J. Sci. Comput.*, 26(1):55–85, 2004.

Distance Problems for Hermitian Matrix Pencils

Shreemayee Bora and Ravi Srivastava

Abstract

A matrix pencil $L(z) = zA - B$ is said to be Hermitian if its coefficient matrices A and B are Hermitian. If $L(z)$ is Hermitian, then a finite real eigenvalue say, λ, of $L(z)$ is said to be of definite type if x^*Ax is non zero for every non zero eigenvector x corresponding to λ. It is said to be of mixed type if it is not of definite type. The type of an eigenvalue at ∞ is the type of 0 as an eigenvalue of $-\mathrm{rev}\, L(z) := zB - A$.

We introduce a definition of eigenvalue type based on the homogeneous form $L(\alpha, \beta) = \alpha A - \beta B$ of the pencil. This is combined with homogeneous Hermitian ϵ-pseudospectrum $\Lambda_\epsilon^{\mathrm{Herm}}(L(\alpha, \beta))$ of the pencil defined by

$$\begin{aligned}\Lambda_\epsilon^{\mathrm{Herm}}(L(\alpha,\beta)) &= \Big\{(\alpha,\beta) \in \mathbb{C}^2 \setminus \{(0,0)\} : \det(\alpha(A+\Delta_1) - \beta(B+\Delta_2)) = 0,\ (\Delta_1)^* = \Delta_1,\\ &\quad (\Delta_2)^* = \Delta_2, \|\Delta_1\|^2 + \|\Delta_2\|^2 < \epsilon^2\Big\},\end{aligned}$$

where $\|\cdot\|$ denotes the 2-norm, to analyze the evolution of real eigenvalues of definite type with respect to Hermitian perturbations. If all eigenvalues of $L(z)$ are real and of definite type, then this analysis leads to a method for computing the distance from $L(z)$ to a nearest Hermitian pencil with at least one eigenvalue of mixed type with respect to the norm $\|L\| := \sqrt{\|A\|^2 + \|B\|^2}$ and a Hermitian pencil that attains this distance.

Definite and definitizable pencils are classes of Hermitian pencils with important applications whose eigenvalues are all real and of definite type. If $L(z)$ is a definite pencil, then it is characterized by the property that its Crawford number $\gamma(A, B) := \min\limits_{\|x\|=1} \sqrt{(x^*Ax)^2 + (x^*Bx)^2}$ is positive. The theory and computation of the Crawford number is well researched (see, for example [4, 2] and [3]) and it is known that it gives the distance from $L(z)$ to a nearest Hermitian pencil that is not definite which respect to the norm specified above. The characterization of definite pencils in terms of the distribution of their eigenvalues based on type as given in [1] implies that a nearest Hermitian pencil that is not definite has an eigenvalue of mixed type. Therefore our results lead to an algorithm for computing the Crawford number and a nearest Hermitian pencil that is not definite. We prove that such a pencil in fact has a defective eigenvalue. The algorithm is based on the bisection principle and if the definite pencil is of size n, then each step of the bisection requires the computation of a largest eigenvalue of a positive definite matrix of size n or a smallest eigenvalue of a negative definite matrix of size n and associated eigenvector(s). Numerical experiments show that the proposed algorithm compares favorably with existing methods for computing the Crawford number. The distance from a definitizable pencil to a nearest Hermitian pencil that is not definitizable may also be computed in a similar way.

This work is a completion of the research presented in Householder Symposium XVIII that resolves the conjectures made earlier.

References

[1] M. Al-Ammari and F. Tisseur. Hermitian matrix polynomials with real eigenvalues of definite type. Part 1: Classification. *Linear Algebra Appl.,* 436(10): 3954–3973, 2012.

[2] Crawford, Charles R. *The numerical solution of the generalised eigenvalue problem.* Thesis (Ph.D.)–University of Michigan. Ann Arbor, 1970.

[3] Higham, Nicholas J. and Tisseur, Françoise and Van Dooren, Paul M. Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems. *Linear Algebra Appl.,* 351/352: 455–474, 2002, Fourth special issue on linear systems and control.

[4] Stewart, G. W. and Sun, Ji Guang. *Matrix perturbation theory.* Computer Science and Scientific Computing, Academic Press Inc., Boston, MA, 1990.

Preconditioning for Low-Rank Matrix Completion via Trust-Regions over one Grassmannian

Nicolas Boumal and P.-A. Absil

Abstract

We present an algorithm for low-rank matrix completion (LRMC) based on Riemannian optimization (or optimization on manifolds) called RTRMC. We first show how, using a least-squares loss function, LRMC amounts to estimating the column space (or the row space, whichever is smallest) of the target matrix. Then, the search space is the space of linear subspaces, which is called the Grassmann manifold. Although abstract in nature, the Grassmann manifold may be endowed with a Riemannian geometry, which means all the now well-established tools from Riemannian optimization can be applied. In particular, RTRMC uses a Riemannian trust-region method. Its global and local convergence analyses carry over gracefully.

Unfortunately, a bad condition number for the target matrix seems to translate into a quadratically worse condition number for the Hessian of the cost function at the solution, which typically slows down convergence dramatically. We show one efficient way to precondition RTRMC which, experimentally, reduces the condition number drastically. Interestingly, preconditioning trust-regions on manifolds amounts to changing the Riemannian metric. The proposed preconditioner can then be seen to be strongly related to the data-induced metrics championed by [5, 6].

RTRMC is described in [3] and [4] without preconditioning. We intend to update the latter reference with preconditioning in the future. For code, visit `http://sites.uclouvain.be/absil/RTRMC/`.

Context

LRMC is the task of recovering a matrix $X \in \mathbb{R}^{m \times n}$ based on knowledge of some of its entries X_{ij}, $(i, j) \in \Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ and knowing that it has rank $r \ll m \leq n$. Noisy versions of LRMC let the knowledge of X_{ij} be tainted by noise and let X only be approximately of rank r.

Assume that we use a least-squares loss function with, for simplicity of exposition, unit weights and no regularization. Then, LRMC can be framed as the optimization problem

$$\min_{\hat{X} \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (\hat{X}_{ij} - X_{ij})^2, \quad \text{such that } \operatorname{rank}(\hat{X}) = r. \tag{1}$$

Convex programming approaches have proven remarkably powerful to solve this problem. They consist for example in replacing the nonconvex rank constraint by a convex, low-rank inducing regularization term such as the nuclear norm (sum of the singular values). Unfortunately, by dropping the rank constraint, the search space becomes much larger than the complexity of the sought object warrants: it increases from a dimension of $r(m + n - r)$ to mn.

Because mining all of $\mathbb{R}^{m \times n}$ in search of an answer may simply be too costly, a number of algorithms have seen the day which make the sacrifice of convexity (and its nice performance guarantees) in exchange for computational efficiency. The typical starting point is the observation that the set of rank-r matrices may be parameterized as $\hat{X} = UW$, with $U \in \mathbb{R}^{m \times r}$ and $W \in \mathbb{R}^{r \times n}$:

$$\min_{U \in \mathbb{R}^{m \times r}, W \in \mathbb{R}^{r \times n}} \sum_{(i,j) \in \Omega} ((UW)_{ij} - X_{ij})^2, \quad \text{such that } U, W \text{ are full-rank.} \tag{2}$$

But factorizations of the form UW are not unique. Indeed, for any invertible matrix M, the pairs (U, W) and $(UM, M^{-1}W)$ are equivalent. Consequently, the search space is still larger than it needs to be and convergence guarantees for optimization methods (which typically assume isolated critical points) may be harder to establish.

Riemannian optimization is perfectly tailored to accommodate such invariances while preserving computational efficiency and convergence guarantees [1]. This motivates numerous investigations which can be classified in two categories. On one side, significant efforts are devoted to the study of Riemannian geometries for the set of fixed-rank matrices [1, 7]. On the other side, some authors, including ourselves, leverage the specific least-squares formulation (at the expense of flexibility in the choice of loss function) to frame LRMC on one or two Grassmannians.

Clearly, fixing either U or W reduces (2) to an easily solved linear least-squares problem. RTRMC replaces W explicitly by W_U, the optimal value for W obtained by fixing U. Regularization (not shown) ensures W_U is well-defined. The product UW_U is only a function of the column space of U, so that $f(U) = \sum_{(i,j)\in\Omega}((UW_U)_{ij} - X_{ij})^2$ descends as a well-defined cost function over the Grassmann manifold and we may restrict our attention to orthonormal U's, for numerical stability.

As indicated above, the Hessian of f may be ill-conditioned close to a solution U if $X \approx UW_U$ is badly conditioned itself (where the conditioning of X is evaluated as σ_1/σ_r and $\sigma_1 \geq \cdots \geq \sigma_m$ are the singular values of X). In an attempt to explain this observation, we simplify the problem and assume all entries (i, j) are observed. Then, if the noise is small, close to convergence at U it can be shown that the Hessian is well approximated by the mapping $H \mapsto H(W_U W_U^T)$ defined on the tangent spaces of the Grassmannian. Since U is orthonormal, it follows that

$$\operatorname{cond}(\operatorname{Hess} f(U)) \approx \operatorname{cond}(W_U W_U^T) = \operatorname{cond}^2(X). \tag{3}$$

This suggests the cheap preconditioner $H \mapsto H(W_U W_U^T)^{-1}$, which is observed to be efficient.

References

[1] P.-A. Absil, L. Amodei, and G. Meyer. Two Newton methods on the manifold of fixed-rank matrices endowed with Riemannian quotient geometries. *Computational Statistics*, 2013.

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 406–414. 2011.

[4] N. Boumal and P.-A. Absil. Low-rank matrix completion via trust-regions on the Grassmann manifold, 2012. Available on Optimization Online.

[5] B. Mishra, K. Adithya Apuroop, and R. Sepulchre. A Riemannian geometry for low-rank matrix completion. *CoRR*, abs/1211.1550, 2012.

[6] T. Ngo and Y. Saad. Scaled gradients on Grassmann manifolds for matrix completion. In *Advances in Neural Information Processing Systems 25*, pages 1421–1429, 2012.

[7] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.

Optimal CUR Matrix Decompositions

Christos Boutsidis and David Woodruff

Abstract

Given as inputs a matrix $A \in R^{m \times n}$ and integers $c < n$ and $r < m$, the CUR factorization of A finds $C \in R^{m \times c}$ with c columns of A, $R \in R^{r \times n}$ with r rows of A, and $U \in R^{c \times r}$ such that $A = CUR + E$. Here, $E = A - CUR$ is the residual error matrix. Compare this to the SVD factorization (let $k < \rho = rank(A)$), $A = U_k S_k V_k' + A_{\rho - k}$. The SVD residual error $A_{\rho-k}$ is the best possible, under some rank constraints. The matrices $U_k \in R^{m \times k}$ and $V_k \in R^{n \times k}$ contain the top k left and right singular vectors of A. In CUR, C and R contain actual columns and rows of A, a property which is desirable for feature selection and data interpretation. This last property makes CUR attractive in a wide range of applications.

From an algorithmic perspective, the challenge is to construct C, U, and R quickly to minimize the approximation error $||A - CUR||_F^2$. The definition below states the precise optimization problem:

The CUR Problem. Given $A \in R^{m \times n}$ of rank $\rho = rank(A)$, rank parameter $k < \rho$, and accuracy parameter $\epsilon > 0$, construct $C \in R^{m \times c}$ with c columns from A, $R \in R^{r \times n}$ with r rows from A, and $U \in R^{c \times r}$, with c, r, and rank(U) being as small as possible, in order to reconstruct A within relative-error :

$$||A - CUR||_F^2 \leq (1 + \epsilon)||A - A_k||_F^2.$$

Here, $A_k = U_k S_k V_k' \in R^{m \times n}$ is the matrix from the best rank k SVD approximation to A.

Despite the significant amount of work and progress on CUR, spanning both the numerical linear algebra community and the theoretical computer science community, there are several important questions which still remain unanswered:

- **Optimal CUR:** Are there any $(1+\epsilon)$-error algorithms selecting the optimal number of columns/rows?
- **Input-sparsity-time CUR:** Are there any input-sparsity-time relative-error algorithms for CUR?
- **Deterministic CUR:** Are there any deterministic, polynomial-time, relative-error CUR algorithms?
- **Rank k CUR:** Are there any relative-error CUR algorithms constructing U with rank at most k?

We settle all of these questions via presenting a randomized, input-sparsity-time and a deterministic, polynomial-time algorithm for the CUR problem. Both algorithms achieve a relative-error bound with $c = O(k/\epsilon)$ columns, $r = O(k/\epsilon)$ rows, and U being a $c \times r$ matrix of rank k. Additionally, a matching lower bound is proven, indicating that both algorithms select the optimal number of columns and rows - up to constant factors - and the optimal $rank(U)$.

To design our optimal algorithms for CUR we combine existing tools on column subset selection such as (i) leverage-score sampling, (ii) BSS sampling (i.e., deterministic sampling similar to the method of Batson, Spielman, and Srivastava, 2009), and (iii) adaptive sampling, with novel tools for input-sparsity-time column subset selection and deterministic versions of adaptive sampling.

Stable Discrete Empirical Interpolation Method based Quadrature Schemes for Nonlinear Model Reduction

Russell L. Carden and Danny C. Sorensen

Abstract

The Discrete Empirical Interpolation Method (DEIM) is a technique for model reduction of nonlinear dynamical systems [1]. It is based upon a modification to proper orthogonal decomposition designed to reduce the computational complexity for evaluating reduced order nonlinear terms. The DEIM approach is based upon an interpolatory projection and only requires evaluation of a few selected components of the original nonlinear term. Although DEIM has been effective on some very difficult problems, it can under certain conditions introduce instabilities in the reduced order model. In particular, if the Jacobian of the original nonlinear term is symmetric definite, then the Jacobian of the corresponding reduced order nonlinear term need not be symmetric nor definite. This can lead to a reduced model that fails to capture the important features of the full model.

We consider a class of nonlinear dynamical systems that can be derived from a variational principle. For such problems there is an associated functional, and a solution to the dynamical system is a critical point of this functional. Rather than apply the DEIM approximation to the nonlinear terms in the dynamical system, we apply the DEIM approximation to the corresponding terms in the functional. The DEIM approximation then determines a quadrature rule for approximating these terms. This approach preserves the symmetry of the Jacobian of the nonlinear terms in the reduced order model, however it does not preserve the defniteness. In order preserve definiteness, the DEIM quadrature rule must be stable. By including more DEIM points than DEIM basis vectors, we show how to construct stable DEIM based quadrature rules. We demonstrate this approach on a problem for which regular DEIM has difficulties. We show how the weighted least squares residual method can be used to generalize this approach to systems that are not derived from a variational principle.

References

[1] S. Chaturntabut and D.C. Sorensen, *Discrete emperical interpolation for nonlinear model reduction*, SIAM J. Sci Comput. 32(5): pp. 2737-2764, 2010.

Improving the Maximum Attainable Accuracy of Communication-Avoiding Krylov Subspace Methods

Erin Carson and James Demmel

Abstract

Krylov subspace methods (KSMs) are a class of iterative algorithms commonly used for solving linear systems $Ax = b$ where A is large and sparse. For simplicity, we restrict our discussion to linear systems where A is real, square, and full rank. In classical KSM implementations, the updates to the solution x_n and updated residual r_n in iteration n consist of one or more sparse matrix-vector multiplications and vector operations. On modern computer architectures, these operations are *communication bound*: the movement of data, rather than computation, is the limiting factor in performance. Recent efforts have thus focused on communication-avoiding Krylov methods (CA-KSMs), based on s-step formulations, which, under certain assumptions on sparsity structure, enable implementations that reduce both parallel and sequential communication cost by a factor of $O(s)$. This can translate into significant speedups in practice [7]. For a thorough overview of CA-KSMs and related work, see [5].

Although equivalent in exact arithmetic, CA-KSMs and their classical counterparts can have significantly different behavior in finite precision; relative to the classical implementation, CA-KSMs can require a greater number of iterations to converge to the same tolerance and/or suffer a decrease in *maximum attainable accuracy*, i.e., the accuracy with which the method can solve $Ax = b$ on a computer with unit round-off ϵ. Well-conditioned polynomial bases such as Newton and Chebyshev can be used to alleviate these effects (see, e.g., [8]), although convergence and accuracy eventually deteriorate for high s. Both theoretical and empirical comparisons with the stability and convergence properties of classical implementations are thus crucial in accessing the practicality of CA-KSMs for real-world problems.

Studies on the finite precision behavior of CA-KSMs have thus far been empirical in nature; in this talk, we review our recent work [1], the first quantitative analysis of the maximum attainable accuracy of finite precision CA-KSMs, specifically communication-avoiding conjugate gradient (CA-CG) and biconjugate gradient (CA-BICG).

In classical KSMs, errors made in updates to x_n and r_n in each iteration can accumulate and cause deviation of the *true residual*, $b - Ax_n$, and the *updated residual*, r_n. Writing the true residual as $b - Ax_n = r_n + (b - Ax_n - r_n)$, we can bound its norm by $\|b - Ax_n\| \leq \|r_n\| + \|b - Ax_n - r_n\|$. When the updated residual r_n is much larger than $b - Ax_n - r_n$, the true residual and the updated residual will be of similar magnitude. However, as $\|r_n\| \to 0$, the size of the true residual depends on $\|b - Ax_n - r_n\|$. If this deviation grows large, it can limit the maximum attainable accuracy. Many have studied the maximum attainable accuracy of classical KSMs; see, e.g., [3, 4, 6, 9, 10].

Residual replacement strategies, in which the updated residual is replaced by the true residual in certain iterations, have been shown to maintain $\|b - Ax_n - r_n\| = O(\epsilon)\|A\|\|x\|$ for classical KSMs [10]. The replacement iterations are chosen based on an estimate of $\|b - Ax_n - r_n\|$ such that (1) the size of the deviation of residuals remains small and (2) convergence of the finite precision Lanczos process is undisturbed [9, 10].

We derive a bound on the deviation of the true and updated residuals in CA-CG and CA-BICG in finite precision. This bound can be iteratively updated within the method inexpensively, i.e., without asymptotically increasing communication or computation costs. Our bound allows insight into how maximum attainable accuracy is affected by s and choice of polynomial basis.

Furthermore, following the derivation in [10] for classical KSMs, our bound enables an implicit residual replacement strategy for maintaining agreement between residuals to within $O(\epsilon)\|A\|\|x\|$ in CA-CG and CA-BICG. Numerical experiments on a small set of test matrices verify that, for cases where the updated residual converges, the residual replacement strategy can enable accuracy of $O(\epsilon)\|A\|\|x\|$ with a small number of residual replacement steps, reflecting improvements of up to 7 orders of magnitude.

Much work remains to be done on the analysis of finite precision CA-KSMs. We plan to extend the analysis and scheme developed here to other CA-KSMs, such as CA-BICGSTAB [2], as well as preconditioned variants. We briefly mention the broader open problem of developing convergence and stability theories for CA-KSMs in finite precision.

References

[1] E. Carson and J. Demmel, *A residual replacement strategy for improving the maximum attainable accuracy of s-step Krylov subspace methods*, SIAM J. Matrix Anal. Appl. (in press).

[2] E. Carson, N. Knight, and J. Demmel, *Avoiding communication in nonsymmetric Lanczos-based Krylov subspace methods*, SIAM J. Sci. Comp., 35 (2013).

[3] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.

[4] M. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 213–229.

[5] M. Hoemmen, *Communication-avoiding Krylov subspace methods*, PhD thesis, EECS Dept., U.C. Berkeley, 2010.

[6] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.

[7] M. Mohiyuddin, M. Hoemmen, J. Demmel, and K. Yelick, *Minimizing communication in sparse matrix solvers*, in Proc. ACM/IEEE Conference on Supercomputing, 2009.

[8] B. Philippe and L. Reichel, *On the generation of Krylov subspace bases*, Appl. Numer. Math, 62 (2012), pp. 1171–1186.

[9] G. Sleijpen and H. van der Vorst, *Reliable updated residuals in hybrid Bi-CG methods*, Computing, 56 (1996), pp. 141–163.

[10] H. Van der Vorst and Q. Ye, *Residual replacement strategies for Krylov subspace iterative methods for the convergence of true residuals*, SIAM J. Sci. Comput., 22 (1999), pp. 835–852.

Structure-Preserving Model Reduction for Nonlinear Port-Hamiltonian Systems

Saifon Chaturantabut, Christopher A. Beattie and Serkan Gugercin

Abstract

Port-Hamiltonian systems can be used for modeling port-based network systems, as well as other physical systems described by the Euler-Lagrange equations. Each of these systems is generally defined by a power conserving geometric structure that captures the basic interconnection laws, together with a Hamiltonian function that describes the total stored energy. We consider the finite-dimensional port-Hamiltonian systems of the form:

$$\begin{aligned} \dot{\mathbf{x}} &= (\mathbf{J} - \mathbf{R})\, \nabla_{\mathbf{x}} H(\mathbf{x}) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y} &= \mathbf{B}^T \nabla_{\mathbf{x}} H(\mathbf{x}), \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the n-dimensional state vector; $H : \mathbb{R}^n \to [0, \infty)$ is the *Hamiltonian* function; $\mathbf{J} = -\mathbf{J}^T \in \mathbb{R}^{n \times n}$ is the *structure matrix* describing the interconnection of energy storage elements in the system; $\mathbf{R} = \mathbf{R}^T \geq \mathbf{0}$ is the $n \times n$ *dissipation matrix* describing energy loss in the system and, $\mathbf{B} \in \mathbb{R}^{n \times m}$ is the *input/output matrix* describing how energy enters and exits the system. It can be shown that the family of the port-Hamiltonian systems given in (1) is always *stable* and *passive*, i.e.

$$H(\mathbf{x}(t_1)) - H(\mathbf{x}(t_0)) \leq \int_{t_0}^{t_1} \mathbf{y}(t)^T \mathbf{u}(t)\, dt, \tag{2}$$

which implies that the change in the internal energy of the system, as measured by H, is bounded by the total work done on the system.

The dimension n of this port-Hamiltonian system (1) can get extremely large as the size of the corresponding network model increases. This can result in an intensive computational cost, especially when the simulation is repeatedly required with different inputs for the same system. Therefore, our goal is to construct a low dimensional port-Hamiltonian structure-preserving reduced system that can give accurate output responses for wide-ranging interested inputs.

We extend the earlier work in [1], which proposed a structure and stability preserving model reduction approach for large-scale input-output nonlinear port-Hamiltonian systems. This approach is based on the Petrov-Galerkin projection with a special modification that allows the resulting reduced system to preserve the original port-Hamiltonian structure. Two techniques for constructing reduced-order bases were considered: Proper Orthogonal Decomposition (POD) and a quasi-optimal $\mathcal{H}_2$ model reduction [3]. Although, the resulting low-dimensional reduced systems constructed from both types of bases were shown to accurately capture the original output behavior, they may still have complexity proportional to the full-order dimension n, particularly in the case of *general nonlinear* Hamiltonian functions. Therefore, these reduced systems may not be able to achieve a significant reduction in actual computation time.

This work incorporates the notion of Discrete Empirical Interpolation Method (DEIM) [2] with the structure preserving model reduction approach from [1] to handle the nonlinearity of port-Hamiltonian systems. Since DEIM can completely destroy the port-Hamiltonian structure, a modification is crucially required before applying it to the nonlinear term. The structure-preserving DEIM approximation is derived through the minimization of the enforced structure-preserving form of the nonlinear term at some specially chosen interpolation components, which are corresponding

to the indices selected by the greedy algorithm as done in [2]. The complexity reduction is therefore a result of the fact that the interpolation projection will only require evaluation of the nonlinear term at these few selected components. Each resulting reduced system can be shown to preserve the port-Hamiltonian structure, and hence it maintains the stability and passivity of the original system (2). This work also derives the corresponding a-priori error bounds of the state variables and outputs by using an application of generalized logarithmic norms for unbounded nonlinear operators. The effectiveness of the proposed approach will be shown through a nonlinear ladder network and a Toda lattice model with exponential interactions.

References

[1] C. A. Beattie and S. Gugercin, *Structure-preserving model reduction for nonlinear port- Hamiltonian systems*, Decision and Control and European Control Conference (CDC- ECC), 50th IEEE Conference on, Dec., pp. 6564-6569, 2011 .

[2] S. Chaturantabut and D. C. Sorensen, *Discrete empirical interpolation for nonlinear model reduction*, SIAM J. Sci. Comput., 32(5): pp. 2737–2764, 2010.

[3] S. Gugercin, A. C. Antoulas, and C. A. Beattie, $\mathcal{H}_2$ *model reduction for large-scale linear dynamical systems*, SIAM J. Matrix Anal. Appl., 30, pp. 609-638, 2008.

Two Methods for Computing the Matrix Sign Function

Jie Chen and Edmond Chow

Abstract

We are interested in computing

$$S = \text{sign}(A)$$

for a large, Hermitian matrix A. The matrix sign function plays a key role in electronic-structure calculations, where A stands for a shifted Hamiltonian. The density matrix is the Heaviside function of A, which is simply S after a scaling and a shift. Computing such a matrix is challenging in terms of both floating point operations and storage. One might recall a related argument that the inverse of a large matrix is rarely needed; in our case, however, the elements of the density matrix are used in self-consistent field iterations and hence computing the whole S is necessary.

We have developed two methods that offer parallelization benefits over the traditional Newton's method and the method of Padé approximations, which requires matrix inversions or a large number of matrix solves. The first method is an improvement of the Newton-Schulz iteration

$$X_{k+1} = f(X_k), \qquad X_0 = A,\ f(x) = \frac{1}{2}x(3 - x^2)$$

that leads to $X_k \to S$ for $\|I - A^2\| < 1$. The appealing quadratic convergence of Newton-Schulz is eclipsed by the slow progress in the initial steps. The central idea of the improvement is to accelerate the initial convergence by using a better fixed-point polynomial f that minimizes the condition number of X_{k+1}. We show that the number of fixed-point iterations for reaching single or double precision accuracy is half of that of Newton-Schulz in general. This method heavily uses matrix-matrix multiplications and the parallelization is relatively easy to optimize.

The second method is also based on matrix-matrix multiplications, specifically, A-multiplies. We use a least squares polynomial g_k with degree not exceeding k to approximate the sign function:

$$g_k(A) \to S \quad \text{as} \quad k \to \infty.$$

Compared with Chebyshev approximations that cannot handle well the singularity of the sign function, we design an L^2 inner product with a special weight function to bypass the singularity and we show a linear convergence of g_k. Preconditioning techniques (such as deflation) can be incorporated to improve the convergence. This method is favorable when the A-vector multiplication has a cost lower than quadratic, such as when A is sparse, is Toeplitz, or is a fully dense kernel matrix. This method can also naturally be extended to computing $\text{sign}(A)b$ for any vector b.

Preconditioned Methods for Sampling Multivariate Gaussian Distributions

Edmond Chow and Yousef Saad

Abstract

In a wide variety of probabilistic simulations, it is necessary to compute sample vectors from a multivariate Gaussian distribution with zero mean and a given covariance matrix. For large-scale problems, this task is computationally expensive, and despite the availability of several approaches, fast methods are still highly desired since such sampling remains a computational bottleneck in these simulations.

For a symmetric positive definite covariance matrix A, the canonical method of constructing a Gaussian sample $y \sim N(0, A)$ is to compute $y = Sz$, where $A = SS^T$ and z is a standard normal vector. With this definition of S, it is clear that the covariance of y will be the matrix A. As a result, any S that satisfies $A = SS^T$ can be used, for example, the lower triangular Cholesky factor or principal square root of A. Use of the Cholesky factor is most common.

Instead of Cholesky factorization, we focus on a category of sampling methods based on matrix polynomials. Here, sample vectors of the form $p(A)z$ are computed, where p is a polynomial chosen such that $p(A)$ approximates the principal square root of A. One need not form the matrix $p(A)$ explicitly. Instead, for each z, $p(A)z$ is computed from a sequence of products with the matrix A. Two main types of polynomial methods have been used: those that choose $p(A)$ as an expansion of Chebyshev or other orthogonal polynomials, and those that construct the sample from a Krylov subspace.

In this talk, we show how these polynomial sampling methods can be preconditioned. Unlike the linear system case, preconditioning here changes the problem being solved. However, a sample with the desired covariance can still be recovered.

Consider a factorization $G^T G$ that approximates A^{-1}, which we may call a preconditioner. Now consider using the preconditioned matrix GAG^T, rather than A, in a polynomial sampling method, to produce a sample

$$\tilde{w} = p(GAG^T)z$$

which is approximately distributed as $N(0, GAG^T)$. Since GAG^T is well-conditioned, we expect that only a small number of terms is required in the polynomial approximation to the square root of GAG^T. To construct a sample with the desired covariance, apply G^{-1} to each sample $\tilde{w}$,

$$\tilde{y} = G^{-1}\tilde{w} = G^{-1}p(GAG^T)z$$

which is approximately distributed as $N(0, A)$. Such a sample approximates $G^{-1}(GAG^T)^{1/2}z$. The matrix $S = G^{-1}(GAG^T)^{1/2}$ satisfies

$$SS^T = G^{-1}(GAG^T)^{1/2}(GAG^T)^{1/2}G^{-T} = A$$

as desired, but it is not a Cholesky factor or square root of A. The idea leading to the method described here is that S can be *any* of an infinite number of quantities that satisfies $SS^T = A$. How well the covariance of $\tilde{y}$ approximates A depends on the accuracy of p and not on G. The convergence rate depends on the quality of the approximation $G^T G \approx A^{-1}$. For example, in the extreme case when we have an exact Cholesky factorization at our disposal, then $G^T G = A^{-1}$, and we only need to take $p(A) = I$, i.e., $\tilde{y}$ becomes $\tilde{y} = G^{-1}p(GAG^T)z = G^{-1}z$. This corresponds

to the standard method based on Cholesky (G^{-1} is the lower triangular Cholesky factor of A). However, we now have the option of using an approximate factorization instead of an exact one.

As usual, the preconditioned matrix GAG^T is not formed explicitly. Instead, G and G^T are applied to vectors in these methods. To construct the desired sample, however, we must be able to easily *solve* with G. (The roles of matrix-vector multiplications and solves are reversed if we define the preconditioner GG^T as an approximation to A.) These requirements are more demanding than the usual requirements for preconditioners.

In this talk, we will also propose a preconditioner satisfying the above requirements and that are very suitable for covariance matrices used to model Gaussian processes. These covariance matrices include those based on piecewise polynomial, exponential, Matérn, and Gaussian radial distribution functions.

To motivate the choice of preconditioner, first consider that the inverse of a covariance matrix, known as a *precision matrix*, measures the conditional independence between data points. For Gaussian Markov distributions, this precision matrix is sparse. For many other types of Gaussian distributions, including those just mentioned, the precision matrix shows a strong decay in the size of its elements. These observations motivate using a preconditioner that is a sparse approximation to the inverse of the covariance matrix, e.g., factorized sparse approximate inverse preconditioners. In our experimental tests, we found that preconditioning with this choice of preconditioners can reduce computation time by at least a factor of 10 on large problems.

A Framework for Regularization via Operator Approximation

Julianne Chung, Misha Kilmer and Dianne O'Leary

Abstract

Large-scale inverse problems arise in many applications such as astronomy, biomedical imaging, surveillance, and nondestructive evaluation. We consider large-scale inverse problems of the form

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m \geq n$, denotes the forward operator, $\mathbf{b} \in \mathbb{R}^m$ represents the observed data, $\mathbf{n} \in \mathbb{R}^m$ is additive noise, and $\mathbf{x} \in \mathbb{R}^n$ is the desired solution. Given $\mathbf{A}$ and $\mathbf{b}$, the goal of the inverse problem is to reconstruct $\mathbf{x}$.

Most inverse problems are ill-posed, meaning small perturbations in the observation may lead to large changes to the solution [5, 8, 10]. To mitigate this difficulty, regularization is often used, adding constraints that suppress the amplification of noise during the inversion process. Choosing an appropriate regularization method and a good regularization parameter to balance fidelity to the model with satisfaction of the constraints is key to solving any inverse problem. Regularization approaches based on spectral filtering can be highly effective [6]. However, these methods require computing the singular value decomposition (SVD) and choosing appropriate regularization parameters, which can be prohibitively expensive for large-scale problems.

In this talk, we propose a framework for regularization that uses operator approximations to determine the regularization for large-scale problems in which the SVD of $\mathbf{A}$ is not available. The framework consists of three steps:

1. Find a related but simpler operator $\widehat{\mathbf{A}} \approx \mathbf{A}$, whose SVD is easily computable.
2. Choose a regularization method and find suitable regularization operators/parameters for the approximate problem,
$$\mathbf{b} = \widehat{\mathbf{A}}\mathbf{x} + \mathbf{n}\,. \tag{2}$$
3. Find the regularized solution of the original problem (1), using the same regularization method, operators/parameters determined in step 2.

One of the key advantages of our proposed framework is that we apply the regularization to the *original* problem (1). The operator approximation is only used to determine the regularization and regularization parameter in a computationally efficient way. Although a solution to the approximate problem (2) may provide an estimate of the desired solution for the original problem (1), previous researchers have observed that the approximate problem can yield poor reconstructions [7, 2]. Instead, we propose to only use the operator approximation in step 2 of our framework, so that sophisticated regularization and regularization parameter selection methods can be utilized for problems for which it is too expensive to apply the methods directly.

A variety of regularization approaches can be incorporated into this framework but we focus here on the recently developed windowed regularization [3], a generalization of Tikhonov regularization in which different regularization parameters are used in different regions of the spectrum. We derive bounds on the perturbation to the computed solution and on the error in the computed solution, resulting from using the regularization determined for the approximate operator. We demonstrate the effectiveness of our method in computations using operator approximations such as sums of Kronecker products [9], BCCB (block circulant with circulant blocks) matrices [1], and Krylov subspace approximations [4].

References

[1] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Comput., 9 (1988), pp. 766–771.

[2] D. Chen, *Numerical Methods for Edge-Preserving Image Restoration*, PhD thesis, Tufts University, 2012.

[3] J. Chung, G. Easley, and D. P. O'Leary, *Windowed spectral regularization of inverse problems*, SIAM J. Sci. Comp., 33 (2011), pp. 3175–3200.

[4] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis, 2 (1965), pp. 205–224.

[5] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, 2010.

[6] P. C. Hansen, J. G. Nagy, and D. P. O'Leary, *Deblurring Images: Matrices, Spectra and Filtering*, SIAM, Philadelphia, PA, 2006.

[7] J. Kamm and J. G. Nagy, *Kronecker product and SVD approximations in image restoration*, Linear Algebra Appl., 284 (1998), pp. 177–192.

[8] J. L. Mueller and S. Siltanen, *Linear and nonlinear inverse problems with practical applications*, vol. 10, SIAM, 2012.

[9] J. G. Nagy, M. K. Ng, and L. Perrone, *Kronecker product approximation for image restoration with reflexive boundary conditions*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 829–841.

[10] C. R. Vogel, *Computational methods for inverse problems*, SIAM, 2002.

Testing Matrix Functions Using Identities

Edvin Deadman and Nicholas J Higham

Abstract

We introduce a new method for testing matrix function algorithms, based on the residuals in functional identities. This work was motivated by a collaboration between the University of Manchester and NAG, a software company specializing in numerical libraries. During this collaboration, many of the latest algorithms for computing matrix functions were implemented for the NAG Library. One of the difficulties encountered was in testing the accuracy and stability of the implementations.

Matrix function algorithms typically involve combinations of scaling, square roots [4], Taylor series and Padé approximants. Backward rounding error analyses are usually not possible and numerical experiments are essential to demonstrate the accuracy and stability of the algorithms.

Let $\hat{Y}$ denote the computed approximation to a matrix function $Y = f(X)$. The normwise relative *forward error* is given by $\|Y-\hat{Y}\|/\|Y\|$. Suppose that there exists a ΔX such that $\hat{Y} = f(X+\Delta X)$. Then the normwise relative *backward error* is the smallest possible value of $\|\Delta X\|/\|X\|$.

The relative condition number $\kappa_f(X)$ is defined as

$$\kappa_f(X) = \lim_{\epsilon\to 0} \sup_{\|E\|\leq\epsilon\|X\|} \frac{\|f(X+E)-f(X)\|}{\epsilon\|f(X)\|}.$$

It gives an upper bound on the effect of small changes in X on the solution Y. A useful rule of thumb [5, §1.6] states that the forward error is approximately bounded by the product of the condition number and the backward error.

When testing an algorithm, two questions must be considered. First, how can we compute the forward or backward errors? Second, how small should we expect them to be?

The standard approach used in the matrix functions literature is motivated by the fact that uncertainties in X (for example due to rounding errors) mean that we cannot reasonably demand backward errors smaller than the unit roundoff, u. In practice, backward errors within a reasonably small factor of u are deemed to be acceptable. However, backward errors themselves are not usually available when testing an algorithm. Instead, the algorithm is judged to be behaving in a stable manner if the forward error is smaller than $\kappa_f(X)u$. If this criterion does not hold, then the backward error must necessarily be greater than u and the algorithm cannot be backward stable. The forward error is obtained by computing the "exact" solution Y by diagonalizing X in high precision arithmetic (a tiny random perturbation ensures that X is diagonalizable [3]).

This method has been used very successfully to test new algorithms. However, it has two limitations.

1. At NAG, software is tested on several different computer architectures. High precision arithmetic is unlikely to be available, so forward errors cannot be computed using the method above. (Even if high precision arithmetic is available, the increased cost of floating point operations may limit the sizes of the test matrices that can be used.)

2. Computing the condition number involves maximizing the perturbation in the data over all directions. Hence $\kappa_f(A)u$ represents a worst-case scenario. Even if the forward error is smaller than $\kappa_f(A)u$ it is still that the backward error for that particular computation is larger than u. Hence a forward error smaller than $\kappa_f(A)u$ merely shows that the algorithm is behaving in a way consistent with backward stability.

We have approached the problem by using the residuals from matrix function identities to test matrix function algorithms. By expanding to linear order in the backward error terms, we are able to generalize the idea of comparing the forward error with $\kappa_f(A)u$ by deriving quantities with which the residuals in the identities should be compared. For identities of the form $f(g(A)) = A$ (such as $e^{\log A} = A$) the relevant quantity is

$$\mathrm{res}_{\max} = u\{1 + \kappa_f(g(A))\}.$$

For identities in which the product $f(A)g(A)$ is known (e.g. $A^{1/3}A^{2/3} = A$) the relevant quantity is

$$\mathrm{res}_{\max} = u\frac{\|A\|}{\|f(A)g(A)\|} \max_{\substack{\|E_1\|\le 1\\ \|E_2\|\le 1}} \|L_f(A, E_1)g(A) + f(A)L_g(A, E_2)\|,$$

where $L_f(A, E_1)$ denotes the Fréchet derivative of f in the direction E_1. We are able to show how, provided that Fréchet derivatives can be computed, this maximum can be reliably estimated in a manner similar to that used to estimate the 1-norm condition number of a matrix [6, §3.4].

This approach obviates the need to use high precision arithmetic to obtain "exact" solutions, solving the first of the limitations above.

To deal with the second limitation, we show how a normwise relative backward error estimate can be obtained directly from the residuals in the matrix function identities. This is done by, again, expanding to linear order in the backward error terms and solving the resulting least squares problem to give a backward error estimate.

A matrix function algorithm may give an inaccurate result because the condition number of the matrix is large. Alternatively, the matrix may be well-conditioned but could have a property that causes difficulties for the algorithm (for example the Schur-Parlett algorithm [2] struggles with certain eigenvalue distributions). Numerical experiments show that our new methods work particularly well in the latter case, successfully distinguishing between algorithms that are performing stably and algorithms that are not (although the methods are unable to attribute the errors to individual matrix function evaluations in the identities). For highly ill-conditioned matrices however, the linearization assumptions made in our analysis may not be valid, and the methods can fail. Thus care is required in the selection of test problems when using these methods in practice.

References

[3] E. B. Davies. *Approximate diagonalization. SIAM J. Matrix Anal. Appl.*, 29(4):1051–1064 (2007)

[2] P. I. Davies and N. J. Higham. *A Schur-Parlett algorithm for computing matrix functions. SIAM J. Matrix Anal. Appl.*, 25(2):464–485 (2003)

[3] E. Deadman and N. J. Higham. *Testing matrix functions using identities. In preparation.*

[4] E. Deadman, N. J. Higham, and R. Ralha. Blocked Schur algorithms for computing the matrix square root. In *Applied Parallel and Scientific Computing: 11th International Conference PARA2012*, Lecture Notes in Comput. Sci. 7782, Springer-Verlag, 171–182 (2013)

[5] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, 2nd edition (2002)

[6] N. J. Higham. *Functions of Matrices: Theory and Computation.* SIAM (2008)

Coupled Matrix/Tensor Decompositions: an Introduction

Laurent Sorber, Mikael Sorensen and Marc Van Barel and Lieven De Lathauwer

Abstract

Decompositions of higher-order tensors are becoming more and more important in signal processing, data analysis, machine learning, scientific computing, optimization and many other fields. A new trend is the study of coupled matrix/tensor decompositions (i.e., decompositions of multiple matrices and/or tensors that are linked in one or several ways). Applications can be found in various fields and include recommender systems, advanced array processing systems, multimodal biomedical data analysis and data completion. We give a short overview and discuss the state-of-the-art in the generalization of results for tensor decompositions to coupled matrix/tensor decompositions. We briefly discuss the remarkable uniqueness properties, which make these decompositions important tools for signal separation. Factor properties (such as orthogonality and triangularity, but also nonnegativity, exponential structure, etc.) may be imposed when useful but are not required for uniqueness per se. Also remarkable, in the exact case the decompositions may under mild conditions be computed using only tools from standard linear algebra. We touch upon the computation of inexact decompositions via numerical optimization (this is discussed in more detail in the companion talk by M. Van Barel). We illustrate some of the ideas using Tensorlab, a Matlab toolbox for tensors and tensor computations that we have recently released, and of which version 2 provides a comprehensive framework for the computation of (possibly constrained) coupled matrix/tensor decompositions.

Communication Avoiding Algorithms for Linear Algebra and Beyond

James Demmel

Abstract

Algorithms have two costs: arithmetic and communication, i.e. moving data between levels of a memory hierarchy or processors over a network. Communication costs (measured in time or energy per operation) already greatly exceed arithmetic costs, and the gap is growing over time following technological trends. Thus our goal is to design algorithms that minimize communication. We present algorithms that attain provable lower bounds on communication, and show large speedups compared to their conventional counterparts. These algorithms are for direct and iterative linear algebra, for dense and sparse matrices, as well as direct n-body simulations. Several of these algorithms exhibit perfect strong scaling, in both time and energy: run time (resp. energy) for a fixed problem size drops proportionally to the number of processors p (resp. is independent of p). Finally, we describe extensions to algorithms involving arbitrary loop nests and array accesses, assuming only that array subscripts are affine functions of the loop indices.

Model Reduction Techniques for Fast Nonlinear Inversion

Eric de Sturler, Serkan Gugercin, Misha Kilmer, Chris Beattie, Saifon Chaturantabut, and Meghan O'Connell

Abstract

As many inverse problems are severely underdetermined, it is appropriate to solve for a medium or image parameterized with a relatively modest number of parameters. Moreover, the parametrization may provide regularization, as it does for our approach, reducing the inverse problem to a nonlinear optimization problem. Nevertheless, the computational burden remains very high, because the nonlinear least squares optimization requires the solution of many large linear systems per iteration. This is expensive, and the ever-increasing number of measurements enabled through advances in engineering increases this computational burden further. We show how techniques from interpolatory model reduction can drastically reduce the cost of inverse problems by reducing the cost of solving many forward problems.

Specifically, we discuss how reduced order transfer functions and their gradients can approximate the objective function and the associated Jacobian with little loss of accuracy but significantly reduced cost. The quality and performance of our method is demonstrated for several synthetic diffuse optical tomography (DOT) problems. An important and intriguing result is that global projection bases for reduced models derived from only a few interpolation points for a single reconstruction can be used, without further expensive linear solves, for very different reconstructions. We will explain this result and discuss interesting related questions in parametric inversion, model reduction, optimization, and linear solvers.

For diffuse optical tomography (DOT), we solve the following partial differential equation with appropriate boundary conditions, posed in the frequency domain,

$$-\nabla \cdot (D(x;p)\nabla u(x;\omega)) + \left(a(x;p) + \frac{\imath\omega}{\nu}\right) u(x;\omega) = s(x;\omega),$$

where x is position, ω is the frequency modulation of light, ν is the speed of light in the medium, $u(x;\omega)$ is the photon flux, $D(x;p)$ and $a(x;p)$ are respectively the diffusion and absorption field depending nonlinearly on the parameter vector p, and $s(x;\omega)$ is an input light source (here a point source on or near the surface) [3]. For a given parameter vector p, giving diffusion and absorption fields, this PDE can be solved for u (the forward problem). The result provides estimates for measurements by a set of detectors. We solve a nonlinear least squares problem for p by minimizing the difference between actual (noisy) measurements and computed measurements, thereby determining the diffusion and absorption fields. We use the PaLS parametrization [1], which models the shapes of anomalies using compactly supported radial basis functions (CSRBFs) and level sets, for $D(x;p)$ and $a(x;p)$, and we solve the nonlinear least squares problem using the TREGS solver proposed in [6].

After discretization of the PDE, the computed measurements for frequency ω and parameter vector p are given by the matrix

$$C\left(\frac{\imath\omega}{\nu}E + A(p)\right)^{-1} B,$$

which one readily recognizes as the transfer function of a discrete dynamical system. Here, E is the identity, except for zeros on the diagonal corresponding to grid points at the sides where sources and detectors are located, $A(p)$ is the discretization of the diffusion and absorption operators, the

columns of B correspond to source locations, and the rows of C correspond to detector locations (or some localized quadrature rule). Hence the evaluation of the objective function for least squares minimization involves the solution of this system for *many* sources (all columns of B) and for a modest number of frequencies ω_j. The computation of the Jacobian (with respect to p) of the objective function can be done using a co-state method, making it roughly as expensive as solving a similar system for all detector locations. The main cost of the optimization is solving a very large number of large, sparse, linear systems (discretized 3D PDEs).

Interpolatory model reduction provides the tools to replace this expensive transfer function by a much smaller (reduced order) transfer function [2] that requires only the solution of small linear systems but nevertheless provides accurate approximations to both the objective function as well as the Jacobian. The reduced transfer function is obtained by computing global projection bases following [4], requiring a modest number of large linear solves. However, after computing those bases once, the optimization can be done without any further solution of large, linear systems. Indeed, in our numerical experiments, the same global bases can be used for the reconstruction of substantially different images in the same 'experimental setting' (discretization, number of CSRBFs, and source and detector locations). Hence, *no further off-line costs are incurred for subsequent, distinct reconstructions.* This is an intriguing and potentially important result, as it suggests that for our method reduction bases can be computed once in an off-line phase and used for many distinct reconstructions. Potential explanations for this result, based on 'inexact reduction bases for interpolatory model reduction' [5] will be discussed. We will also discuss alternative approaches, such as a method proposed in [7], which has some nice theoretical results, but requires the computation of new local reduction bases at every optimization step.

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grants No. NSF-DMS 1025327, NSF DMS 1217156 and 1217161, NSF-DMS 0645347, and NIH R01-CA154774.

References

[1] A. Aghasi, E. Miller, and M. E. Kilmer. Parametric level set methods for inverse problems. *SIAM Journal on Imaging Science*, 4:618–650, 2011.

[2] A. Antoulas, C. Beattie, and S. Gugercin. Interpolatory model reduction of large-scale dynamical systems. In J. Mohammadpour and K. Grigoriadis, editors, *Efficient Modeling and Control of Large-Scale Systems*, pages 2–58. Springer-Verlag, 2010.

[3] S. R. Arridge. Optical tomography in medical imaging. *Inverse Problems*, Vol. 16:R41–R93, 1999.

[4] U. Baur, P. Benner, C. Beattie, and S. Gugercin. Interpolatory projection methods for parameterized model reduction. *SIAM Journal on Scientific Computing*, 33:2489–2518, 2011.

[5] C. Beattie, S. Gugercin, and S. Wyatt. Inexact solves in interpolatory model reduction. *Linear Algebra and its Applications*, 2011. Appeared on-line at, doi:10.1016/j.laa.2011.07.015.

[6] E. de Sturler and M. E. Kilmer. A regularized Gauss-Newton trust region approach to imaging in diffuse optical tomography. *SIAM Journal on Scientific Computing*, 33:3057 – 3086, 2011.

[7] V. Druskin, V. Simoncini, and M. Zaslavsky. Solution of the time-domain inverse resistivity problem in the model reduction framework Part I. one-dimensional problem with SISO data. *SIAM Journal on Scientific Computing*, 35(3):A1621 – A1640, 2013.

Backward Error and Conditioning of Fiedler Companion Linearizations.

Fernando De Terán and Françoise Tisseur

Abstract

Given a matrix polynomial $P(\lambda) = \sum_{i=0}^{k} \lambda^i A_i$, with $A_i \in \mathbb{C}^{n\times n}$, the *Polynomial Eigenvalue Problem* (PEP) consists of finding pairs (λ_0, v), with $v \neq 0$, such that $P(\lambda_0)v = 0$. Such pairs are known as *eigenpairs*, $\lambda_0 \in \mathbb{C}$ is a *finite eigenvalue* and $v \in \mathbb{C}^{n\times n}$ is an *eigenvector* associated to the eigenvalue λ_0. *Infinite eigenvalues* of $P(\lambda)$ are defined as the zero eigenvalues of the *reversal* polynomial $\mathrm{rev}P(\lambda) = \lambda^k P(1/k)$. Classically, the PEP has been solved by linearizing the polynomial $P(\lambda)$, namely, by transforming the PEP into a *Generalized Eigenvalue Problem* (GEP) and then applying the QZ algorithm to this GEP. More precisely, a *linearization* of $P(\lambda)$ is a matrix polynomial of degree one (or *matrix pencil*), $L(\lambda)$, such that $E(\lambda)L(\lambda)F(\lambda) = \mathrm{diag}(P(\lambda), I_{(k-1)n})$, for some matrix polynomials $E(\lambda), F(\lambda)$ with constant nonzero determinant. Every linearization of $P(\lambda)$ has the same eigenvalues as $P(\lambda)$, though not the same eigenvectors. Nonetheless, in the classical linearizations (the *Frobenius companion linearizations*) the eigenvectors of $P(\lambda)$ can be easily recovered from the ones of the linearization. The Frobenius companion linearizations are the ones used in MATLAB's routine `polyeig` to solve the PEP. However, the Frobenius linearizations present several drawbacks. For instance, they do not preserve any of the structures that matrix polynomials arising in applied problems usually have, like symmetry, palindromicity, etc. The introduction of new families of linearizations opens the possibility of using other linearizations for the PEP. We are particularly interested in the *Fiedler companion linearizations*, which include the Frobenius companion linearizations. Fiedler pencils are particular cases of *companion forms*, namely, templates which are easily constructible from the coefficients A_i of $P(\lambda)$, without performing any arithmetic operation. These companion forms present the following desirable properties:

(a) They are always linearizations, regardless of $P(\lambda)$ being regular or singular (though, in this talk, we focus on *regular* polynomials).

(b) The left and right eigenvectors of $P(\lambda)$ are easily recovered from those of the companion form.

(c) The leading coefficient of the companion form is block diagonal, thereby reducing the computational cost of the Hessenberg-triangular reduction step of the QZ algorithm.

(d) They can be easily transformed into a block upper-triangular form revealing zero and infinite eigenvalues, if any.

(e) They are the source of structured linearizations.

In this talk, we analyze some numerical features of Fiedler companion forms. When the PEP is solved by linearization, it is transformed into a GEP, with different conditioning and backward error than the PEP. We will first display explicit formulas for the eigenvectors of any Fiedler pencil $F(\lambda)$ which, in particular, will show that the eigenvectors of $P(\lambda)$ can be easily recovered from those of the Fiedler pencil. Using these formulas, we will compare the backward error of approximate eigenpairs of $F(\lambda)$ with the backward error of the corresponding eigenpair of $P(\lambda)$, and we obtain bounds for the ratio between these two backward errors. We also compare the conditioning of eigenvalues in $F(\lambda)$ and $P(\lambda)$, and derive bounds for the corresponding ratio as well. We will see, in particular, that if the matrix polynomial is well scaled (i.e., $\|A_i\|_2 \approx 1$, for all $i = 0, 1, \ldots, k$), then the Fielder companion linearizations have good conditioning and backward stability properties. Some numerical experiments will be shown to illustrate our theoretical results.

Parallel Asynchronous Matrix Factorization for Large-Scale Data Analysis

Inderjit S. Dhillon, H. Yun, C.J. Hsieh, H.F. Yu and S.V.N. Vishwanathan

Abstract

In many applications in large-data analysis, the data to be analyzed is not fully known upfront, but arrives incrementally. For example, ratings of movies by users are being constantly added, and a social graph keeps evolving. When the data is in the form of a matrix, this corresponds to matrix entries arriving incrementally. Typical analysis techniques on such data first form "latent factor" models, which correspond to matrix factorization models, for example, low-rank matrix completion, non-negative matrix factorization, etc. As the data sets grow in size, there is a critical need to parallelize the corresponding matrix factorization techniques. Most existing efforts, as in traditional numerical linear algebra, alternate between computation and communication, requiring explicit synchronization. In this talk, I will discuss asynchronous parallel algorithms for matrix factorization techniques used in data analysis. We name our framework NOMAD as portions of the factors migrate *asynchronously* from processor to processor depending on where these portions are needed. These algorithms are well-suited to the case where data arrives incrementally and the factorization needs to be continually updated. As an example, we implement a parallel, asynchronous low-rank matrix completion algorithm and test it on a HugeWiki data set that has about 50 million rows, 40 thousand columns and 2.7 billion non-zeros. Our asynchronous algorithm involves fine-grain parallelism, but still is able to achieve impressive speedups on traditional supercomputing hardware (the Stampede cluster at the Texas Advanced Computing Center) as well as on commodity hardware (Amazon Web Services), and outperforms competing methods. If time permits, I will discuss a recently initiated effort to try and ease the burden on the programmer for rapidly prototyping and testing a class of such algorithms.

Changes of Canonical Structure Information of Matrix Pencils associated with Generalized State-space Systems.

Andrii Dmytryshyn, Stefan Johansson and Bo Kågström

Abstract

We consider *generalized state-space* (or *descriptor*) systems

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where $A, E \in \mathbb{C}^{n\times n}$ and E is non-singular, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, $D \in \mathbb{C}^{p\times m}$, and $x(t), y(t), u(t)$ are the *state, output,* and *input (control)* vectors, respectively. The system characteristics of (1) are typically ill-posed problems, i.e., small perturbations in the matrices can lead to drastic changes in the system characteristics. Such problems can be represented and analyzed from the *canonical structure information* (elementary divisors, column and row minimal indices) of the block-structured *system pencil*:

$$\mathcal{S}(\lambda) := \begin{bmatrix} A & B \\ C & D \end{bmatrix} - \lambda \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix}, \quad \det(E) \neq 0. \tag{2}$$

Two state-space pencils $\mathcal{S}'(\lambda)$ and $\mathcal{S}(\lambda)$ are called *feedback-injection equivalent* if and only if there exist non-singular matrices

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} S_{11} & 0 \\ S_{21} & S_{22} \end{bmatrix}, \tag{3}$$

such that $\mathcal{S}'(\lambda) = R\mathcal{S}(\lambda)S$.

$\mathcal{S}(\lambda)$ can also be considered under strict equivalence. Then for any non-singular matrices P and Q the pencil $\mathcal{S}'(\lambda) = P\mathcal{S}(\lambda)Q$ does not need to be of the form (2). Nevertheless, if it is of the form (2) then there exist R and S of the form (3) such that $\mathcal{S}'(\lambda) = R\mathcal{S}(\lambda)S$. So two state-space pencils are feedback-injection equivalent if and only if they are strictly equivalent. In particular, it means that any system pencil (2) has the same canonical structure information under strict and feedback-injection equivalence, respectively. The canonical structure information (one may also think about canonical forms here) depends discontinuously on the entries of the matrices involved.

Using *versal deformations* (i.e., a simple form to which all pencils $\mathcal{B}(\lambda)$ close to $\mathcal{S}(\lambda)$ can be reduced by the considered transformations that smoothly depend on the entries of $\mathcal{B}(\lambda)$), we prove that there exist an arbitrarily small dense perturbation $\mathcal{W}(\lambda)$ (zeros in the λ-part can be perturbed too), and non-singular P and Q such that

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \left(\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} W_1 & W_3 \\ W_2 & W_4 \end{bmatrix} - \lambda \left(\begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} W_5 & W_6 \\ W_7 & W_8 \end{bmatrix} \right) \right) \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \mathcal{S}'(\lambda)$$

if and only if there exist an arbitrarily small perturbation $\mathcal{V}(\lambda)$ of the form (2) (zeros in the λ-part are fixed and are not allowed to be perturbed) and non-singular R and S of the form (3) such that

$$\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \left(\begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} V_1 & V_3 \\ V_2 & V_4 \end{bmatrix} - \lambda \left(\begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} V_5 & 0 \\ 0 & 0 \end{bmatrix} \right) \right) \begin{bmatrix} S_{11} & 0 \\ S_{21} & S_{22} \end{bmatrix} = \mathcal{S}'(\lambda).$$

Note that the sufficiency is obvious.

This reduces the problem of describing the changes of the canonical structure information under feedback-injection equivalence to the problem of describing these changes under strict equivalence which is known (e.g., see [1, 2, 3]). We also explain how the closest neighbours (cover relations) in the closure hierarchy can be obtained.

We present our results by constructing *stratifications* (i.e., closure hierarchy graphs) [4] of system pencils (2). The stratifications show the possible changes of the canonical structure information caused by arbitrarily small perturbations of the original system pencil $\mathcal{S}(\lambda)$.

Altogether, it appears that the stratification graph Δ of $\mathcal{S}(\lambda)$ is an induced subgraph of the stratification graph Γ of $\mathcal{S}(\lambda)$ considered as a general matrix pencil (i.e., it is a subset of the vertices of a graph Γ together with any edges whose endpoints are both in this subset).

We also present the special case of the system (1) with no direct feedforward (i.e., the matrix D is zero).

References

[1] A. Edelman, E. Elmroth, and B. Kågström. A Geometric Approach to Perturbation Theory of Matrices and Matrix Pencils. Part I: Versal Deformations. *SIAM J. Matrix Anal. Appl.*, 18(3):653–692, 1997.

[2] A. Edelman, E. Elmroth, and B. Kågström. A Geometric Approach To Perturbation Theory of Matrices and Matrix Pencils. Part II: A Stratification-Enhanced Staircase Algorithm. *SIAM J. Matrix Anal. Appl.*, 20:667–699, 1999.

[3] E. Elmroth, S. Johansson, and B. Kågström. Stratification of controllability and observability pairs — Theory and use in applications. *SIAM J. Matrix Anal. Appl.*, 31(2):203–226, 2009.

[4] B. Kågström, S. Johansson, and P. Johansson. StratiGraph Tool: Matrix Stratification in Control Applications. In L. Biegler, S. L. Campbell, and V. Mehrmann, editors, *Control and Optimization with Differential-Algebraic Constraints*, chapter 5. SIAM Publications, 2012. ISBN 978-1-611972-24-5.

The inverse complex eigenvector problem for real tridiagonal matrices

Beresford Parlett, Froilán M. Dopico and Carla Ferreira

Abstract

A classic open problem in Numerical Linear Algebra is "the nonsymmetric tridiagonal eigenvalue problem". The statement of this problem is very simple: to find a backward stable algorithm to compute all eigenvalues (and, optionally, all eigenvectors) of an $n \times n$ nonsymmetric tridiagonal matrix with a cost of $O(n^2)$ flops. While it is very simple to state, the solution of this problem is very hard and the main reason of this difficulty is that backward stable orthogonal similarities, like those used in the eigenvalue QR algorithm, do not preserve the nonsymmetric tridiagonal structure and, so, lead to algorithms with $O(n^3)$ computational cost. This has motivated to explore other options and algorithms of *dqds*-type are perhaps the most natural approach, since they take full advantage of the tridiagonal structure and have been very successful to solve the symmetric tridiagonal eigenvalue problem and the bidiagonal singular value problem in a backward stable way and with $O(n^2)$ cost. Several versions of these algorithms have been developed in the last years by Beresford Parlett and Carla Ferreira. They behave extremely well in practice in terms of fast performance, exhibit an $O(n^2)$ computational cost, and they are much faster than other algorithms available in the literature for "the nonsymmetric tridiagonal eigenvalue problem". However, very rarely, these non-symmetric *dqds*-algorithms fail to compute backward stable outputs, and to find a shift-strategy which always guarantees backward stability is still an open problem. In this context, a sensible idea is to refine simultaneously the eigenvalues computed by dqds and the eigenvectors obtained from them to get eigenpairs (eigenvalue/right eigenvector) or eigentriples (eigenvalue/right eigenvector/left eigenvector) with tiny residuals which guarantee that the computed eigenpairs/eigentriples are the exact ones of a nearby *tridiagonal* matrix, i.e., we would get backward structured stability. Again, this is very easy to say, but not easy to do, and one of the reasons that makes this difficult is that some aspects of "the nonsymmetric tridiagonal eigenvalue problem" are not well understood. In this talk, we investigate one of such aspects.

We will focus on an $n \times n$ *real* nonsymmetric tridiagonal matrix T. It is well known that T may have complex eigenvalues. Imagine that we have computed a complex eigentriple $(\hat{\lambda}, \hat{\mathbf{x}}, \hat{\mathbf{y}}^*)$ of T and we want to determine if $(\hat{\lambda}, \hat{\mathbf{x}}, \hat{\mathbf{y}}^*)$ is an exact eigentriple of a nearby real nonsymmetric tridiagonal matrix $T+E$. It is easy to see that this leads to $4n$ real equations for the $3n-2$ unknown real entries of E, which will be inconsistent unless the vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}^*$ are very particular. In plain words we are saying that not every complex triple $(\hat{\lambda}, \hat{\mathbf{x}}, \hat{\mathbf{y}}^*)$ is an eigentriple of a real nonsymmetric tridiagonal matrix. The situation for the condensed representations of real *unreduced* tridiagonal matrices in terms of $2n-1$ real parameters which are used in *dqds*-algorithms is even more complicated, since for them most complex pairs $(\hat{\lambda}, \hat{\mathbf{x}})$ cannot be eigenpairs of real nonsymmetric tridiagonals represented in these ways unless the vector $\hat{\mathbf{x}}$ is very particular.

The purpose of this talk is to present simple necessary and sufficient conditions to determine if a given complex triple $(\hat{\lambda}, \hat{\mathbf{x}}, \hat{\mathbf{y}}^*)$ is an eigentriple of a real nonsymmetric tridiagonal matrix and if a complex pair $(\hat{\lambda}, \hat{\mathbf{x}})$ is an eigenpair of a condensed representation of a real unreduced tridiagonal matrix. These conditions can be checked with $O(n)$ cost and, in addition, we will provide simple and efficient algorithms with $O(n)$ cost for computing the corresponding tridiagonal matrices. So, our results can be used for computing structured backward errors and for discarding (and potentially improving) computed eigentriples/eigenpairs which cannot be solutions of the original problem.

Identifying Influential Entries in a Matrix

Petros Drineas and Abhisek Kundu

Abstract

Modern data sets are often represented by large matrices, since a matrix $A \in \mathbb{R}^{m \times n}$ provides a natural structure for encoding information about m objects, each of which is represented with respect to n features. Examples include the bag-of-words representation of document corpora (where the objects are documents and the features are words), microarray data (where the objects are genomes and the features are genes), stocks and their temporal resolution, etc. In each of these application areas, practitioners spend vast amounts of time analyzing the data in order to understand, interpret, and ultimately use this data for some application-specific task.

We seek to identify a small set of *influential entries* for a matrix A, in the sense that these entries capture most of the "information" in A. In order to have a concrete optimization objective in mind, we will seek a small set of entries of A that can be used to accurately *reconstruct* A. So, let $\Omega = \{(i_t, j_t), t = 1 \dots s\}$ be a set of s index pairs, indicating which s entries of A are chosen. Then a *reconstruction algorithm* takes as input the entries of A that correspond to the index pairs in Ω and outputs a matrix $\tilde{A}$ such that $\tilde{A}$ is close to A, in the sense that $A - \tilde{A}$ is small (as measured by some matrix norm).

We briefly expand on potential practical uses of identifying influential entries of a matrix. *First,* the selected entries can be used for exploratory data analysis tasks: the disproportionate influence that they exert on the matrix is a starting point for further probing the object-feature combinations that the entries represent. As a matter of fact, our approach is ideal for such tasks: it is based on the *element-wise leverage scores*, essentially a probability distribution over the entries of A. Spikes of that distribution (entries of high leverage) represent object-feature combinations that correspond to disproportionately important data elements. Practitioners would be interested to investigate whether these entries are the result of measurement noise (and thus should be removed as outliers) or whether they are the manifestation of an underlying structural property of the data generating process. *Second,* the selected entries form a succinct representation of the input matrix and can be used to perform common machine learning tasks, such as clustering and classification, instead of the full matrix. This succinct representation could result in faster computations (especially in tasks that take advantage of matrix sparsity) and could potentially improve accuracy by removing less informative pieces of the data, essentially *regularizing* the task at hand.

Two lines of research have investigated the selection of influential entries from a matrix in order to accurately reconstruct it. Achlioptas and McSherry pioneered the selection of entries from a matrix with probabilities that depend on their magnitude squared and proved that applying, for example, the Singular Value Decomposition on the selected entries (with the remaining entries of the matrix set to zero) to get a low-rank approximation, results in a matrix $\tilde{A}$ that is provably close to A. The resulting reconstruction error (with respect to common unitarily invariant norms such as the spectral and Frobenius norms) scales as a function of the Frobenius norm of A, which, while useful if one needs only a coarse approximation to the actual input in the context of streaming applications, might be prohibitively large in fine-grain exploratory data analysis applications.

A second line of research focused on the *matrix completion* problem. Motivated by recommendation systems and collaborative filtering, a large body of work focused on the matrix completion problem. Existing work has almost exclusively focused on the case where the user has no control over the selection of the matrix entries, but instead entries are revealed uniformly at random. Assuming that

the input matrix has rank ρ and satisfies the so-called *incoherence assumptions*, prior work argued that given a uniform sample Ω consisting of $O\left((m+n)\,\rho \ln(m+n)\right)$ elements of A, the solution to the following nuclear norm minimization problem recovers A exactly, with high probability:

$$\begin{aligned} \min_{\Phi \in \mathbb{R}^{m\times n}} \quad & \|\Phi\|_* \\ \text{subject to} \quad & \Phi_{ij} = A_{ij}, \quad \text{for all } (i,j) \in \Omega \end{aligned} \tag{1}$$

The incoherence assumptions basically guarantee that no entry of A exerts disproportionate influence over the matrix and thus a uniform sample is an accurate sketch of the matrix. As we will see below, our approach identifies influential entries and then uses the framework of matrix completion to argue that these entries can be used to reconstruct the input matrix.

Our results

Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank ρ, with $\rho \ll \min\{m,n\}$, and let the Singular Value Decomposition (SVD) of A be $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times\rho}$ is the matrix of the left singular vectors of A, $\Sigma \in \mathbb{R}^{\rho\times\rho}$ is the diagonal matrix of singular values of A, and $V \in \mathbb{R}^{n\times\rho}$ is the matrix of the right singular vectors of A. We can now define an element-wise probability distribution on the entries of A (the *element-wise leverage scores*) as follows.

Definition 1. *Given a matrix $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$ of rank $\rho \ll \min\{m,n\}$, the **element-wise leverage scores** (for all i,j) are*

$$p_{ij} = \frac{1}{2}\frac{\left\|U_{(i)}\right\|_2^2 + \left\|V_{(j)}\right\|_2^2 - \left\|U_{(i)}\right\|_2^2\left\|V_{(j)}\right\|_2^2}{(m+n)\,\rho - \rho^2} + \frac{1}{4}\frac{(UV^T)_{ij}^2}{\rho} + \frac{1}{4}\frac{\left|(UV^T)_{ij}\right|}{\sum_{i,j=1}^{n}\left|(UV^T)_{ij}\right|}. \tag{2}$$

$U_{(i)}$ (respectively $V_{(j)}$) denotes the i-th row of U (respectively j-th row of V) as a row vector.

Obviously, $p_{ij} \geq 0$ for all i,j; it is also easy to prove that $\sum_{i,j=1}^{m,n} p_{ij} = 1$. Thus, we can use a simple, element-wise sampling algorithm to sample a set of entries from the matrix A in independent identically distributed trials, with replacement, with respect to the p_{ij}'s. Our main theorem below states that if we sample a sufficient number of entries of A (see eqn. (3) for the precise sample size) with respect to the element-wise leverage scores, then the nuclear norm minimization problem of eqn. (1) recovers the matrix A with constant probability. To the best of our knowledge, our bounds are the best known for matrix completion under no incoherence assumptions, and are only a factor of ρ away from being optimal.

Theorem 1. *Given a matrix $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$ of rank $\rho \ll \min\{m,n\}$, sample s entries of A using the **element-wise leverage scores** of eqn. (2) as the sampling probabilities in s independent, identically distributing trials, with*

$$s > \max\left\{360\,(m+n)\,\rho^2 - 360\rho^3, 42\,(m+n)\,\rho \ln 20\,(m+n)\right\}, \tag{3}$$

to construct a set of element-wise samples from A. Then, with probability at least 0.8, the nuclear norm minimization problem of eqn. (1) returns A as the unique minimizer.

We will also present connections of the element-wise leverage scores to so-called *leverage scores* of rows and columns of matrices and their use in Randomized Numerical Linear Algebra algorithms, as well their use in diagnostic data analysis applications[1].

[1] See `http://arxiv.org/abs/1310.3556` for more details.

A new Framework for Polynomial Filtering in Implicitly Restarted Arnoldi type Algorithms

Zvonimir Bujanović and Zlatko Drmač

Abstract

We explore the theoretical framework of the implicitly restarted Arnoldi [3] and the Krylov-Schur [4] algorithm for computing only a small number of eigenpairs of $A \in \mathbb{C}^{n\times n}$ that satisfy some property (e.g. all eigenvalues contained in a given domain $\Omega \subset \mathbb{C}$, or certain number of eigenvalues closest to the origin or to the imaginary axis etc.). In particular, we focus on the theoretical possibility for implicit restarting of the Krylov-Schur algorithm with arbitrary polynomial filter – a functionality lacking in the original algorithm. Our results reveal a very interesting latent connection between implicit restarting with arbitrary polynomial filter and the partial pole placement (eigenvalue assignment) problem.

The key observation is as follows: Let the Krylov decomposition $AX = XG + xg^*$ be partitioned as

$$A \begin{pmatrix} X_{[1]} & X_{[2]} \end{pmatrix} = \begin{pmatrix} X_{[1]} & X_{[2]} \end{pmatrix} \begin{pmatrix} G_{[11]} & G_{[12]} \\ 0 & G_{[22]} \end{pmatrix} + x \begin{pmatrix} g_{[1]}^T & g_{[2]}^T \end{pmatrix},$$

where $A \in \mathbb{C}^{n\times n}$, $X_{[1]} \in \mathbb{C}^{n\times k}$, $G_{[11]} \in \mathbb{C}^{k\times k}$, $g_{[1]} \in \mathbb{C}^k$, and the columns of $X = \begin{pmatrix} X_{[1]} & X_{[2]} \end{pmatrix} \in \mathbb{C}^{n\times m}$, $m = k + \ell$, span a Krylov subspace $\mathcal{K}_m(A, v)$ which is not A-invariant.

Then, $AX_{[1]} = X_{[1]}G_{[11]} + xg_{[1]}^T$ is an implicitly restarted Krylov decomposition with the subspace $\mathrm{Im}(X_{[1]}) = \mathcal{K}_k(A, \prod_{i=1}^{\ell}(A - \sigma_i I)v)$ if and only if $\sigma_1, \ldots, \sigma_\ell$ are the eigenvalues of $G_{[22]}$.

It is shown that restarting with arbitrary polynomial filter is possible by reassigning some of the eigenvalues of the Rayleigh quotient through a rank-one correction, implemented using only the elementary transformations (translation and similarity) of the Krylov decomposition. This framework includes the implicitly restarted Arnoldi algorithm (IRA), and the Krylov-Schur algorithm with implicit harmonic restart as special cases. Further, it reveals that the IRA algorithm can be turned (without any change in the algorithm) into an eigenvalue assignment method. Since eigenvalue assignment is notoriously ill-conditioned procedure, this revealed connection opens many interesting numerical issues related to implicit restarting. We will discuss some of them.

The new framework is used to tackle the problems of implicit restarting of the second order Arnoldi algorithm (SOAR) [1], which is challenging task with many open problems.

References

[1] Zh. Bai, Y. Su, SOAR: A second-order Arnoldi method for the solution of the quadratic eigenvalue problem, SIAM J. Matrix Anal. Appl. 26 (2005), pp. 640–659.

[2] Z. Bujanović, Z. Drmač, A new framework for implicit restarting of the Krylov-Schur algorithm. Num. Lin. Alg. Appl. 2013. (in review)

[3] D. C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[4] G. W. Stewart, G. W., A Krylov–Schur algorithm for large eigenproblems, SIAM J. Matrix Anal. Appl. 23 (2001), pp. 601–614.

Matrix Functions and Their Krylov Approximations for Large Scale Wave Propagation in Unbounded Domains.

Vladimir Druskin, Alexander Mamonov, Rob Remis and Mikhail Zaslavsky

Abstract

We target large scale hyperbolic problems in unbounded domains. Until now, the method of choice for such problems is explicit finite-difference time domain (FDTD) method coupled with PML absorbing boundary condition, which is hindered by the CFL limit. Instead, we want to apply the Krylov and rational Krylov subspace technique in the matrix function framework, that previously showed great success for diffusion dominated PDEs and model reduction applications.

We consider wave problems in unbounded domains requiring computations of actions of functions of selfadjoint nonnegative PDE operators with *continuous* spectrum on compactly supported initial conditions, e.g.,

$$A^{-1/2}\sin(\sqrt{A}t)b,$$

where A and b are respectively the operator and the initial condition.

To avoid spurious resonances, the reduced order model should *preserve spectral continuity* of the original problem, i.e., it should not contain poles on the main Riemann sheet. That can be achieved by using (non-Hermitian) complex symmetric discretized operators A_N dumped by the perfectly matched layers and so-called stability-corrected time-domain exponential (SCTDE) matrix function

$$\Im\left(A_N^{-1/2}e^{\mathrm{i}\sqrt{A_N}t}\right)b_N,$$

where b_N is the discretized initial condition [1].

First, we approximate the SCTDE using the Krylov subspace projection algorithm, based on the re-normalized Lanczos method. With the same cost per step as the FDTD, it over-performs the latter for large propagation times.

However, convergence of the Krylov subspace approximation of the SCTDE matrix function decelerates due to appearance of *the square root singularity*. The convergence can be improved by employing extended and rational Krylov subspaces, but they become too expensive for the problems of the desired size.

To circumvent the above problem, we suggest to compute the rational Krylov subspace (RKS) implicitly, by partitioning the computational domain and expanding the Schur complements of the subdomain resolvents via *local* RKSs. By proper projecting the A_N onto these RKSs, we obtain a reduced size well-conditioned structured matrix H. The preprocessing stage is quite costly, but can be done "embarrassingly parallel". Matrix-vector multiplications with H are also well suited for high performance computing. Such an approach allows us to exactly reproduce the *global* RKS and accurately compute the corresponding reduced order model on the partition skeleton with significantly less cost.

References

[1] V. Druskin and R. Remis, *A Krylov stability-corrected coordinate-stretching method to simulate wave propagation in unbounded domains*, SIAM J. Sci. Comput., 35 (2013), pp. 313–357.

The Solution of Least-Squares Problems using Preconditioned LSQR

Iain Duff and Mario Arioli

Abstract

We consider the solution of the problem

$$\min_x ||b - Ax||_2^2 \tag{1}$$

where A is a sparse matrix of rank n and order $m \times n$ with $m \geq n$.

It is not easy to obtain a good preconditioner for solving (1). In particular, the use of an incomplete Cholesky decomposition on the normal equations usually requires so much fill-in to be effective that it is equivalent to using a direct method.

We study a class of preconditioners for the augmented system formulation

$$\begin{bmatrix} I_m & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \tag{2}$$

where r is the residual vector equal to $b - Ax$.

We obtain our preconditioner by partitioning the rows of A into basic and non-basic parts and note that the reduced and preconditioned system

$$\begin{bmatrix} I_{m-n} & NB^{-1} \\ B^{-T}N^T & -I_n \end{bmatrix} \begin{bmatrix} r_N \\ Bx \end{bmatrix} = \begin{bmatrix} b_N \\ -b_B \end{bmatrix}, \tag{3}$$

is a symmetric quasi-definite (SQD) matrix so we can use the generalized LSQR method [1, 3] to solve it. We extend the theoretical eigenanalysis for our preconditioned system using the CS decomposition to show that this SQD system is usually well-conditioned and to emphasize that this has absolutely no relationship to the conditioning of the matrix A. The matrix in equation (3) has a symmetric spectrum with the smallest eigenvalues of absolute value 1 and largest eigenvalues equal to $\sqrt{1 + ||NB^{-1}||_2^2}$. It has already been observed [2] that this means that LSQR will require half the number of iterations of MINRES which we verify in our numerical experiments.

We also show the importance of the selection of the basis matrix B in providing a good preconditioner. We show this on small artificial examples where we demonstrate that it is sometimes not possible to obtain a good B even for well-conditioned matrices A. We then develop a two-pass algorithm that uses sparse direct methods for identifying B using threshold rook pivoting with a very high threshold value before separately factorizing the matrix B using the HSL routine MA48 with a normal default value of threshold. We show, in Table 1, data from using a rook pivoting

id	m	n	nnz	$\kappa(A)$	$\sqrt{1 + ||NB^{-1}||_2^2}$		
					u=.1	u=.5	u=1.0
well1033	1033	320	4732	$1.6\ 10^2$	$1.2\ 10^2$	$2.4\ 10^1$	$1.3\ 10^1$
illc1033	1033	320	4719	$1.8\ 10^4$	$7.5\ 10^1$	$2.4\ 10^1$	$1.4\ 10^1$
well1850	1850	712	8755	$1.1\ 10^2$	$8.3\ 10^2$	$4.7\ 10^1$	$2.4\ 10^1$
illc1850	1850	712	8636	$1.4\ 10^4$	$2.0\ 10^3$	$2.7\ 10^1$	$2.1\ 10^1$

Table 1: Results from using rook pivoting to precondition SQD matrices for Paige and Saunders examples.

Test problem class pref

(m,n)	Tree height		$\|\|NB^{-1}\|\|_\infty$		$\sqrt{1+\|\|NB^{-1}\|\|_2^2}$		$\kappa(A)$
(m,n)	SPTs	SPTl	SPTs	SPTl	SPTs	SPTl	
(9975, 5000)	6	8	9.0	14.0	46.3	65.3	1597.1
(19965,10000)	6	9	10.0	16.0	56.8	141.2	2275.4

Test problem class smallw

(m,n)	Tree height		$\|\|NB^{-1}\|\|_\infty$		$\sqrt{1+\|\|NB^{-1}\|\|_2^2}$		$\kappa(A)$
(m,n)	SPTs	SPTl	SPTs	SPTl	SPTs	SPTl	
(10499, 5000)	21	34	42.0	67.0	55.9	91.0	325.0
(20982,10000)	23	35	44.0	69.0	62.4	144.5	465.4

Test problem class kleinberg

(m,n)	Tree height		$\|\|NB^{-1}\|\|_\infty$		$\sqrt{1+\|\|NB^{-1}\|\|_2^2}$		$\kappa(A)$
(m,n)	SPTs	SPTl	SPTs	SPTl	SPTs	SPTl	
(14906, 5000)	12	15	24.0	27.0	73.8	89.1	344.7
(29625,10000)	14	17	28.0	33.0	109.1	111.0	565.7

Table 2: SPTs: root giving tree of least height; SPTl: root for which we have the longest path to a node.

preconditioner on the iconic test matrices from Paige and Saunders.

A common and important class of matrices A are totally unimodular matrices arising in network problems and mixed finite-element methods. The matrix A is the incidence matrix of an undirected graph $\mathcal{G}$ with m edges and n vertices (we assume that $m > n$). Its entries are $-1, 0, 1$ and in each row there are two nonzero entries corresponding to the two nodes identifying an edge. Such matrices have rank $n - 1$ and removal of any column will give a full rank matrix which is still totally unimodular. Fixing one of the nodes as a *root*, we then identify the basis matrix B by generating the corresponding spanning tree for the graph. The spanning tree will depend on the chosen root (column). The entries of a row of NB^{-1} are bounded by the path length to and from root of the two entries in the row of B, so we can reduce the infinity-norm of NB^{-1} by choosing the B matrix associated with the spanning tree with smallest height. The 2-norm of NB^{-1} will be bounded by twice the tree height multiplied by $\sqrt{m}$. We show, in Table 2, the height of trees, the infinity-norm of NB^{-1}, and the condition number for the preconditioned SQD matrix for some test problems generated from the set in [4]. We do this for both the tree of least height (found exhaustively) and for the tree of greatest height. The results support our theory and show that the preconditioned matrix has far better conditioning than the original system. For efficiency, and for computations on larger matrices, we do not do an exhaustive search but find the least height tree using as roots the nodes from the ten densest columns. We have found almost no difference between doing this and doing an exhaustive search.

We will report on experiments that we are currently conducting on larger matrices both for general least-squares problems and those with A an incidence matrix.

References

[1] M. Arioli and D. Orban, *Iterative methods for symmetric quasi-definite linear systems part I: Theory*, Tech. Rep. Cahier du GERAD G-2013-32, GERAD, 2013.

[2] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2011.

[3] M. A. Saunders, *Solution of sparse rectangular systems using LSQR and CRAIG*, BIT, 35 (1995), pp. 588–604.

[4] A. Taylor and D. J. Higham, *CONTEST: A Controllable Test Matrix Toolbox for MATLAB*, ACM Transactions on Mathematical Software, 35 (2009), No. 4, pp. 1–16.

On the Convergence Curves that can be generated by Restarted GMRES

Jurjen Duintjer Tebbens and Gérard Meurant

Abstract

The GMRES method [7], in particular when it is restarted after a small number of iterations, is a very frequently used iterative solver for linear systems with nonsingular, non-Hermitian input matrices. Although its mathematical definition is based on a simple minimization property, convergence analysis is difficult in the sense that there do not seem to exist, in general, clearly defined characteristics of a given linear system that govern the convergence behavior of the method. With normal system matrices it is in the first place the eigenvalue distribution, and secondarily also the projection of the right-hand side vector in the eigenvector basis, that characterize convergence behavior. As we will explain below, this is not the case for general non-normal input matrices. Among a large variety of proposed approaches for analysis of convergence one can mention approaches based on the pseudo-spectrum [9], the field of values [3] or the numerical polynomial hull [4]. Convergence analysis of *restarted* GMRES also considers non-stagnation conditions, see, e.g., [8].

One approach to enhance insight in GMRES convergence behavior is to study matrices with initial vectors that generate the same residual norm history. This was first done in a 1994 paper by Greenbaum and Strakoš [6], where one can find several ways to construct the set of all linear systems with the same convergence curve. Surprisingly, the matrices of the set can have any nonzero spectrum. Together with the fact that any nonincreasing convergence curve is possible [5], this resulted in parametrizations of the set of linear systems generating prescribed residual norms and having prescribed eigenvalues [1]. In [2] it was shown that these parametrizations allow the additional prescription of Ritz values, that is of the spectra of the Hessenberg matrices generated in all subsequent iterations. This implies that the eigenvalue approximations in Arnoldi's method for eigenproblems can be arbitrarily far from the spectrum of the input matrix.

This type of analysis has been applied to the restarted GMRES method, which is more relevant for practice, in a 2011 paper by Vecharinsky and Langou [10]. Let GMRES(m) denote GMRES restarted after every m iterations with the current residual vector and assume that in every restart cycle, all residual norms are equal except for the very last residual norm of each cycle, which is strictly less than the previous residual norm. If we denote the last residual norm of the kth cycle with $\|r_m^{(k)}\|$, then Vecharinsky and Langou showed that any decreasing convergence curve $\|r_m^{(1)}\| > \|r_m^{(2)}\| > \cdots > \|r_m^{(N)}\|$ is possible with any nonzero spectrum of the input matrix, where it is assumed that the total number of inner GMRES iterations Nm is smaller than the system size.

Our work can be seen as an extension of the results in [10]. In the talk we would show that it is possible to prescribe all the residual norms generated by N cycles of GMRES(m), including the norms *inside* cycles, under two conditions: The last residual norm of every cycle does not stagnate, i.e. $\|r_m^{(k)}\| > \|r_{m-1}^{(k)}\|$ for $k = 1, 2, \ldots, N$ and Nm is smaller than the system size. We show that this is possible with any nonzero eigenvalues of the system matrix. In addition, *all* Ritz values generated inside all cycles, can take arbitrary nonzero values.

If we allow stagnation at the end of cycles, the admissible residual norms and Ritz values satisfy an additional, interesting restriction. It is known that stagnation in the GMRES process takes place at and only at iterations where the corresponding Hessenberg matrix is singular, that is, if and only if there is a zero Ritz value. If at some cycle k, $\|r_{m-j}^{(k)}\| = \cdots = \|r_m^{(k)}\|$ or, equivalently, the last j iterations generate a zero Ritz value, then we show that there must be stagnation (or a zero

Ritz value) during the *first* j iterations of the $(k+1)$st cycle. In other words, stagnation at the end of one cycle is literally mirrored at the beginning of the next cycle and residual norms cannot be prescribed at the beginning of that cycle.

The mentioned results are based on a construction of linear systems which generate, when the restarted Arnoldi orthogonalization process is applied and leaving aside the case of stagnation at the end of cycles, subsequent size $(m+1) \times m$ Hessenberg matrices whose entries can be fully prescribed. It turns out that the constructed linear systems have a fascinating property: Besides the fact that they generate Nm prescribed residual norms of GMRES(m), they generate exactly the same Nm residual norms when full GMRES is applied. Moreover, this property can be used to modify the linear systems such, that GMRES(m) converges faster than GMRES($m+i$) for some positive integers i. It thus offers some insight in the causes of the very counterintuitive behavior observed sometimes in practice, where convergence speed decreases when the restart parameter m is increased.

If time allows it, our talk would also comment on some consequences of our results for restarted Arnoldi methods to solve non-Hermitian eigenproblems.

References

[1] M. Arioli, V. Pták, and Z. Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT, 38 (1998), pp. 636–643.

[2] J. Duintjer Tebbens and G. Meurant, *Any Ritz value behavior is possible for Arnoldi and for GMRES*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 958–978.

[3] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.

[4] A. Greenbaum, *Generalizations of the field of values useful in the study of polynomial functions of a matrix*, Linear Algebra Appl., 347 (2002), pp. 233–249.

[5] A. Greenbaum, V. Pták, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.

[6] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent advances in iterative methods, vol. 60 of IMA Vol. Math. Appl., Springer, New York, 1994, pp. 95–118.

[7] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[8] V. Simoncini and D. B. Szyld, *New conditions for non-stagnation of minimal residual methods*, Numer. Math., 109 (2008), pp. 477–487.

[9] L. N. Trefethen and M. Embree, *Spectra and pseudospectra*, Princeton University Press, Princeton, NJ, 2005.

[10] E. Vecharynski and J. Langou, *Any admissible cycle-convergence behavior is possible for restarted GMRES at its initial cycles*, Num. Lin. Algebr. Appl., 18 (2011), pp. 499–511.

Julia: A Fresh Approach to Technical Computing

Jeff Bezanson, Alan Edelman, Stefan Karpinski, Viral Shah and the greater community

Abstract

Julia is a brand new programming language. That sure sounds dull. But something exciting is happening here. Julia was released only two years ago, and user groups have sprouted up in cities across the United States and around the world. A few classes at MIT used Julia last year, and now it is starting to be taught in universities around the world. Julia was unheard of two years ago, and now it is starting to become widely known around the world. Users of other technical languages may feel comfortable. We have often observed that MATLAB users love MATLAB, while R-users are, well frankly, R-users.

Julia is appealing because it not only lets you get the programming job done, but it lets you get it done happily. It gives you the power of expression that you did not know you even wanted or needed, but once given a taste, there is no return. You get performance and parallelism, because they were not afterthoughts. The proper abstractions provide for flexibility, performance, and power that only users of low level languages thought they could have.

To explain what Julia is, we need to explain the world of computing as it looks today at most universities and research institutions. There are the "computer scientists" and the "computational scientists." The "computer scientist" writes low level code, perhaps in C or C++. He or she expects performance from the machine. These programmers tend to be fairly skilled at software development and the best at software abstractions. The "computational scientist" is the catch-all term to describe scientists, engineers, financial analysts, optimization experts, and "big data" analysts who might be using MATLAB or R, for example. These computational scientists are more interested in the human time it takes to build a program. Abstractions show up some times, but mostly it's about getting the job done quickly. They may know programs could run faster in C, but they are not interested in the time it would take to write such a program. Summing up the tradeoff, the computational scientist writes a technical model in a high level language preferring quickly at the expense of a slower computer execution, while the computer scientist might prefer the opposite tradeoff.

Julia has given the world superpowers in many ways. For starters, it brings the world of computational science and computer science together. Developers of technical computing now get performance at a fraction of the productivity cost. Programmers can get further performance by adding processors that are available to them anywhere on the internet. Truly transformational is the IJulia experience, built on the Ipython notebook, which allows scientific communication in a brand new way. Julia webservers allows researchers to publish their models on the web, and anyone in the world can use, test, and verify the results.

The Julia research experience has not followed the traditional – professor keeps it close until publication – method of research. Rather, Julia is more than open source, it is optimistically open source. All of Julia is available to everyone at any time, there is no need to ask anyone permission to run it, use it, or become a trusted developer. The Julia community is friendly, helpful, welcoming and this has added to the Julia experience.

In this talk we demonstrate Julia and describe its promise for numerical linear algebra.

Computing Fréchet Derivatives in Partial Least Squares Regression

Lars Eldén

Abstract

Partial least squares is a common technique for multivariate regression. The procedure is recursive and in each step basis vectors are computed for the explaining variables and the solution vectors. A linear model is fitted by projection onto the span of the basis vectors. The procedure is mathematically equivalent to Golub-Kahan bidiagonalization, which is a Krylov method, and which is equivalent to a pair of matrix factorizations. The vectors of regression coefficients and prediction are non-linear functions of the right hand side. An algorithm for computing the Fréchet derivatives of these functions is derived, based on perturbation theory for the matrix factorizations. From the Fréchet derivative of the prediction vector one can compute the number of degrees of freedom, which can be used as a stopping criterion for the recursion. A few numerical examples are given.

Efficient Solution of Stochastic Partial Differential Equations Using Reduced-Order Models

Howard C. Elman, Virginia Forstall and Qifeng Liao

Abstract

Stochastic partial differential equations arise when components of models such as boundary conditions or operator coefficients are not known with certainty but instead are specified as random fields. There is considerable interest in solving discrete versions of such equations in order to obtain statistical properties such as moments or cumulative distribution functions of the solutions. However, the number or sizes of the associated algebra problems needed for solution may be large, which may make the cost of using such approaches prohibitive.

We discuss two connected ways to address this:

1. *Reduced-order models* [2]. Let the linear algebra problems of interest be denoted $A_{\boldsymbol{\xi}}\mathbf{u} = \mathbf{f}$, where $\boldsymbol{\xi}$ is a vector of parameters associated with the random process. The solution $\mathbf{u} = \mathbf{u}(\boldsymbol{\xi})$ is also a random field depending on $\boldsymbol{\xi}$. We show that in many cases, it is possible to generate a small set of realizations $\{\mathbf{u}(\boldsymbol{\xi}^{(1)}), \mathbf{u}(\boldsymbol{\xi}^{(2)}), \ldots, \mathbf{u}(\boldsymbol{\xi}^{(n)})\}$, so-called snapshots, such that solutions for other parameters can be well represented in the space of snapshots, i.e., $\mathbf{u}(\boldsymbol{\xi}) \approx \hat{\mathbf{u}}(\boldsymbol{\xi}) \equiv Q\boldsymbol{\eta}(\boldsymbol{\xi})$, where Q is a matrix whose columns form a basis for $span\{\mathbf{u}(\boldsymbol{\xi}_j)\}$. One way to find such approximate solutions is to impose a Galerkin condition,

$$Q^T(\mathbf{f} - A_{\boldsymbol{\xi}}\hat{\mathbf{u}}) = 0,$$

   which requires the solution of a *reduced problem* with coefficient matrix $Q^T A_{\boldsymbol{\xi}} Q$. We give examples from models of diffusion and incompressible fluid dynamics in which sufficient accuracy can be obtained with $n \ll N$, i.e., the size of reduced problem is much smaller than N, the size of the discrete model. In these cases, we demonstrate that it is possible to produce accurate statistical analyses at significantly reduced cost [1].

2. *Iterative Solution of Reduced-Order Models.* The approach discussed above depends on the idea that if a reduced problem that represents the dynamics of large-scale discrete systems can be identified, then it will be inexpensive to work with the reduced model. However, when the parameter space (length of $\boldsymbol{\xi}$) is large it may be that a reduced model is significantly smaller than its full-scale counterpart but not so small as to be easily solved by standard (direct algebraic) methods. We show that in this scenario, iterative methods based on multigrid can be applied to the reduced model to reduce costs and enable the efficient solution of moderate-sized reduced-order models.

[1] H. C. Elman and Q. Liao, Reduced basis collocation methods for partial differential equations with random coefficients, *SIAM/ASA J. Uncertainty Quantification* 1:192-217, 2013.

[2] M. A. Grepl, Y. Maday, N. C. Nguyen and A. T. Patera, Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations, M2AN:575-605, 2007.

The Life Cycle of an Eigenvalue Problem

Mark Embree, Jeffrey Hokanson and Charles Puelz

Abstract

We solve eigenvalue problems to understand physical systems, such as the resonant modes of a structure, or band-gaps in a crystal, or the instabilities of flowing fluid. The transition from the motivating physical problem to numerically computed eigenvalues usually requires a series of approximations:

- physical system → mathematical model;
- mathematical model → linear(ized) differential/integral operator;
- differential/integral operator → large discretization matrix;
- large discretization matrix → smaller projected matrix;
- smaller projected matrix → computed eigenvalues (ideally with high relative accuracy).

While the numerical linear algebra community often focuses on the last two steps ("Given A, find eigenvalues"), in this talk we will argue through a handful of concrete examples that a broader view of this process can improve our approach to eigenvalue computations.

First we address the transition from physical system (where eigenvalues correspond to (complex) frequencies of vibration in measured data) to a linear operator. Data collected from an experimental monochord in our laboratory allows us to explore the accuracy of different mathematical models of damping [3]. These models lead to operators with very different asymptotic spectral behavior, which plays a crucial role in inverse spectral theory and affects the convergence of iterative eigensolvers. However, experimental data often determines these larger eigenvalues poorly, due both to frequency (imaginary part), which limits the amount of energy typical initial conditions have in these modes, and damping (real part), which causes what energy is in these modes to decay quickly. When computing, we can calibrate our target accuracy by the degree to which the motivating application determines the eigenvalue.

Discretization itself raises many interesting challenges that can inform our approach to eigenvalue computations. For example, some discretizations cause *spectral pollution*, where spurious eigenvalues within the limits of the essential spectrum. A shift-invert eigenvalue computation applied to such a discretization can quickly find these erroneous eigenvalues. This problem can be avoided, as shown by Davies and Plum [2], by applying the shift-invert transformation *before* discretization, turning the true interior of the spectrum to the exterior of the transformed problem, where no spectral pollution occurs.

Usually we only seek a small subset of the spectrum of the discretization matrix. In contrast, we present an example from the theory of quasicrystals that demands all eigenvalues of a large matrix. The spectrum of the Fibonacci model is a Cantor set whose fractal dimension bounds the rate at which wave-packet solutions to the Schrödinger equation spread (see, e.g., [1]). To approximate this Cantor set, one approximates the quasiperiodic potential with periodic potentials of increasingly long period. The spectrum of such approximations (comprising the union of real intervals) can be found by computing *all* eigenvalues of two (large) symmetric tridiagonal matrices

plus perturbations in the off-diagonal corners. We will show how such eigenvalues can be found expediently, giving insight into the Cantor spectrum [4].

How does discretization affect that performance of iterative eigensolvers? Typically the convergence rate applied to discretized differential operators slows as the discretization is refined. This behavior can be understood clearly at the operator level, where one often cannot build a Krylov subspace $\text{span}\{v, Av, \ldots A^{k-1}v\}$ because Av is not in the domain of the operator A. Such domain considerations can be handled more readily in shift-invert mode, where by applying the Krylov method directly to the infinite-dimensional operator, one can distinguish the convergence rate governed by the operator itself from the effects of discretization.

Finally, we shall briefly address the significant concern of high relative accuracy eigenvalue computations. Discretizations of unbounded operators grow rapidly in norm as the discretization is refined; the difference between a computed eigenvalue and the desired eigenvalue of the operator depends on both the discretization error and the accuracy of the eigensolver, which typically degrades like $\varepsilon_{\text{mach}} n\|A\|$ for conventional algorithms. This perspective can inform the choice of discretization, where a higher-order method (with less sparsity in A) whose eigenvalues converge more rapidly might be favored over a method whose slower convergence rate demands a larger value of $n\|A\|$.

[1] D. Damanik, M. Embree, and A. Gorodetski. "Spectral properties of Schrödinger operators arising in the study of quasicrystals." Rice University CAAM Department Report TR 12-21, 2012.

[2] E. B. Davies and M. Plum. "Spectral pollution." *IMA J. Num. Anal.* 24 (2004) 417–438.

[3] J. M. Hokanson. *Numerically Stable and Statistically Efficient Algorithms for Large Scale Exponential Fitting*, Ph.D thesis, Rice University, in preparation.

[4] C. Puelz and M. Embree. "Computing spectra of Jacobi operators with long periods." In preparation.

On complex J-symmetric eigenproblems

Peter Benner, Heike Faßbender, and Chao Yang

Abstract

In the context of solving the Bethe-Salpeter equations numerically the eigenvalue problem $H_S x = \lambda x$ for complex matrices

$$H_S = \begin{bmatrix} A & -D^H \\ D & -A^T \end{bmatrix} \in \mathbb{C}^{2n\times 2n}, \quad A = A^H, D = D^T \in \mathbb{C}^{n\times n}$$

arises. Typically, n is fairly large, e.g., in the hundreds of thousands, and one is interested in about 25% of the eigenvalues and the corresponding eigenvectors. The matrices H_S belong to the slightly more general class of matrices H_C,

$$H_C = \begin{bmatrix} A & C \\ D & -A^T \end{bmatrix} \in \mathbb{C}^{2n\times 2n}, \quad A, C = C^T, D = D^T \in \mathbb{C}^{n\times n}. \tag{1}$$

Please note, that here X^T denotes transposition, $Y = X^T, y_{ij} = x_{ji}$, no matter whether X is real or complex, while X^H denotes conjugate transposition, $Y = X^H, y_{ij} = \overline{x_{ji}}$.

For

$$J_n = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \quad I_n \in \mathbb{R}^{n\times n}$$

we have

$$J^T = -J = J^{-1}.$$

Whenever the dimension of J_n is clear from the context, we leave of the subscript for ease of notation. Clearly, it holds

$$(H_C J)^T = H_C J.$$

That is, $H_C J$ is complex-symmetric. It is easy to see that $(H_C^2 J)^T = -H^2 J$, that is, $H_C^2 J$ is complex-skewsymmetric.

Matrices like H_C have been called *complex J-symmetrics* in [2], while H_C^2 is called a *complex J-skew-symmetric* matrix. They form a Lie algebra, and a Jordan algebra, resp., the *complex symplectics* are the associated automorphism group, the associated skew-symmetric bilinear form is $< x, y >= x^T J y$. In [2], these (and other) sets of structured matrices are analyzed further. In particular, results concerning structured square root, structured matrix sign decomposition, structured polar decomposition and structured SVD(-like) decomposition are discussed. The resulting matrix of a similarity transformation $X^{-1} H_C X$ is complex J-symmetric, if X is complex symplectic ($X^T J X = J$). For more on structure preserving transformations see [1] where suitable Givens-like, Householder-like and Gauss-like transformations are discussed.

This resembles the situation in the real case, where any matrix of the form

$$H = \begin{bmatrix} A & C \\ D & -A^T \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \quad A, C = C^T, D = D^T \in \mathbb{R}^{n\times n}$$

is Hamiltonian, that is

$$(HJ)^T = HJ.$$

H^2 is skew-Hamiltonian, that is $(H^2J)^T = -(H^2J)$. Any matrix $S \in \mathbb{R}^{2n\times 2n}$ with $S^TJS = J$ is called symplectic. The Hamiltonian matrices form a Lie algebra,the skew-Hamiltonian matrices a Jordan algebra. The associated automorphism group is the group of symplectic matrices, the associated scalar product $< x, y >= x^TJy$. Symplectic similarity transformations preserve the Hamiltonian structure: $S^{-1}HS$ is symplectic again. Eigenvalues of Hamiltonian matrices always occur in pairs $\{\lambda, -\lambda\}$ if λ is real or purely imaginary, or in quadruples $\{\lambda, -\lambda, \overline{\lambda}, -\overline{\lambda}\}$ otherwise. Hence, the spectrum of any Hamiltonian matrix is symmetric with respect to the real and imaginary axis.

The eigenvalues of H_C also display a symmetry [2]: they appear in pairs $(\lambda, -\lambda)$. If x is the right eigenvector corresponding to λ, $H_Cx = \lambda x$, than Jx is the left eigenvector corresponding to the eigenvalue $-\lambda$ of H_C, $(Jx)^TH_C = -\lambda(Jx)$.

In [3] the Jordan form for complex J-symmetric matrices has been derived. As already observed, all nonzero eigenvalues come in pairs $(\lambda, -\lambda)$. The Jordan block associated with $-\lambda$ is just minus the transpose of the Jordan block associated with λ. Moreover, Jordan blocks associated with the eigenvalue zero either are of even size or appear in pairs.

Any complex J-symmetric matrix X is said to be in structured Schur form if

$$X = \begin{bmatrix} R & B \\ 0 & -R^T \end{bmatrix}, \qquad R, B = B^T \in \mathbb{C}^{n\times n}, \tag{2}$$

where the nonzero eigenvalues of R either have positive real part or zero real part and positive imaginary part. We will prove that for any complex J-symmetric matrix H_C there exists a complex symplectic and unitary matrix $W \in \mathbb{C}^{2n\times 2n}$

$$W^TJW = J \qquad W^HW = I,$$

such that W^HH_CW is in structured Schur form (2).

The only algorithm we could find in the literature specifically designed for complex J-symmetric matrices is a Jacobi-like algorithm for computing the structured Schur form [4].

The most popular way to compute the standard Schur form of a general matrix is the QR algorithm. It is tempting to derive a structured QR algorithm for transforming H_C iteratively into structured Schur form. We will discuss why this is not possible and suggest other methods to compute eigenvalues and eigenvectors of H_C. Moreover, we will discuss the additional structure in H_S and how/whether this can be used in eigenvalue computations.

References

[1] D.S. Mackey, N. Mackey, and F. Tisseur. Structured tools for structured matrices. *Electronic Journal of Linear Algebra (ELA)*, 10:106–145, 2003.

[2] D.S. Mackey, N. Mackey, and F. Tisseur. Structured factorizations on scalar product spaces. *SIAM Journal of Matrix Analysis and Applications*, 27(3):821–850, 2006.

[3] C. Mehl. On classification of normal matrices in indefinite inner product spaces. *Electronic Journal of Linear Algebra (ELA)*, 15:84–106, 2006.

[4] C. Mehl. On asymptotic convergence of nonsymmetric Jacobi algorithms. *SIAM Journal of Matrix Analysis and Applications*, 30:291–311, 2008.

New Algorithms for Calculating the H_∞-norm and the Real Stability Radius

Melina A. Freitag, Alastair Spence and Paul Van Dooren

Abstract

The H_∞-norm of a transfer function matrix is an important property for measuring robust stability in classical control theory. We introduce a new fast method for calculating this quantity by extending an algorithm recently introduced in [5].

Consider the continuous time linear dynamical system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times p}$, $C \in \mathbb{C}^{m\times n}$ and $D \in \mathbb{C}^{m\times p}$.

We will focus on continuous time systems, but the idea and method extends to discrete time problems (as was shown for the distance to instability in the discrete case in [10, Chapter 3]). Moreover, the method extends to linear time-invariant descriptor systems.

Let A be stable, then the H_∞-norm of the transfer matrix $G(s) = C(sI - A)^{-1}B + D$ is defined as

$$\|G\|_\infty := \sup_{\omega\in\mathbb{R}} \sigma_{\max}(G(i\omega)), \tag{2}$$

where $\sigma_{\max}$ denotes the maximum singular value of the matrix. Note that for $B = C = I$ and $D = 0$ the reciprocal of $\|G\|_\infty$ is referred to as the distance to instability [11, 4]. The H_∞-norm is used in several applications, for example, in robust control or as an error measure for model order reduction. In this talk we introduce a new quadratically convergent method of estimating the H_∞-norm for the continuous time linear dynamical system given by (1). This method builds on two foundations. First, we use the relation between the singular values of a transfer function matrix and the eigenvalues of a certain Hamiltonian matrix introduced in [4]. Second, we use the implicit determinant method which has its roots in bifurcation analysis and has recently been applied to the problem of finding the stability radius of continuous linear systems [5] (and later extended to discrete systems [10, Chapter 3]).

The standard and most well-known method to compute the H_∞-norm is the Boyd-Balakrishnan-Bruinsma-Steinbuch algorithm [2, 3] which requires repeated computation of all the eigenvalues of a Hamiltonian matrix. A new method to compute the H_∞-norm was recently suggested in [8] and [1]. Both approaches are generalisations of the method in [9], the former uses spectral value sets as generalisations to the matrix pseudospectrum and fast approximations to the spectral value set abscissa, the latter uses structured pseudospectra and extends the idea to descriptor systems.

We propose an algorithm for computing the H_∞-norm based on the implicit determinant method, discuss its implementation (using the staircase reduction) and give numerical examples that illustrate the performance of the method in comparison with other recently developed methods.

If time permits and if there is interest we can show how the above algorithm can be extended to compute the real stability radius of a matrix [6].

References

[1] P. Benner and M. Voigt, *A structured pseudospectral method for h_∞-norm computation of large-scale descriptor systems*, MPI Magdeburg Preprint, (2012).

[2] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its* $\mathbf{L}_\infty$*-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[3] N. A. Bruinsma and M. Steinbuch, *A fast algorithm to compute the* H_∞*-norm of a transfer function matrix*, Systems Control Lett., 14 (1990), pp. 287–293.

[4] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[5] M. A. Freitag and A. Spence, *A Newton-based method for the calculation of the distance to instability*, Linear Algebra Appl., 435 (2011), pp. 3189 – 3205.

[6] ——, *A new approach for calculating the real stability radius*, 2013. submitted.

[7] M. A. Freitag, A. Spence, and P. Van Dooren, *Calculating the* H_∞*-norm using the implicit determinant method*, 2013. submitted.

[8] N. Guglielmi, M. Gürbüzbalaban, and M. L. Overton, *Fast approximation of the* H_∞ *norm via optimization over spectral value sets*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 709–737.

[9] N. Guglielmi and M. L. Overton, *Fast algorithms for the approximation of the pseudospectral abscissa and pseudospectral radius of a matrix*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1166–1192.

[10] M. Gürbüzbalaban, *Theory and methods for problems arising in robust stability, optimization and quantization*, PhD thesis, Courant Institute of Math. Sciences, NYU, 2012.

[11] C. F. Van Loan, *How near is a stable matrix to an unstable matrix?*, in Linear algebra and its role in systems theory (Brunswick, Maine, 1984), vol. 47 of Contemp. Math., Amer. Math. Soc., Providence, RI, 1985, pp. 465–478.

Convergence of restarted Krylov subspace methods for matrix functions

Andreas Frommer, Stefan Güttel, Marcel Schweitzer

Abstract

When approximating $f(A)\mathbf{b}$, the action of a matrix function on a vector, by a Krylov subspace method, the maximum possible number of iterations is often dictated by storage limitations which do not allow to store the full Arnoldi basis, or by the growing computational complexity of evaluating f on a Hessenberg matrix of growing size. Therefore, it may sometimes be impossible to perform the number of iterations needed to reach a prescribed accuracy. To overcome these problems, a number of restart approaches, similar in spirit to restarted methods for linear systems, have been proposed in the literature in recent years [2, 3, 6, 8, 9] and there has been substantial algorithmic advancement concerning stability and computational efficiency. However, a question which remains largely unanswered is under which circumstances convergence of the restarted method can be guaranteed. So far, the only convergence results either cover the case of entire functions or restart length $m = 1$, [1, 3].

In this poster we consider the class of Stieltjes functions, see [7], and a related class, which contain important functions like the (inverse) square root and the matrix logarithm. The general form of a restarted Krylov subspace method obtains successive approximations $\mathbf{f}^{(k)}$ to $f(A)b$ as

$$\mathbf{f}^{(k+1)} = \mathbf{f}^{(k)} + \mathbf{e}_m^{(k)},$$

where $\mathbf{e}_m^{(k)}$ is a Krylov subspace approximation to $e_m^{(k)}(A)b^{(k)}$, $e_m^{(k)}$ being the restart function for the k-th cycle, i.e. we have

$$f(A)b = \mathbf{f}^{(k)} + e_m^{(k)}(A)b^{(k)} \text{ for all } k, \;\; \mathbf{f}^{(0)} = 0,\, b^{(0)} = b,\, e_m^{(0)} = f.$$

Usually, the Krylov subspace approximation is taken to be the Arnoldi approximation, i.e. with the customary representation of the Arnoldi process in the k-th cycle as

$$AV_m^{(k)} = V_m^{(k)} H_m^{(k)} + h_{m+1,m}^{(k)} v_{m+1}^k \hat{e}_m^T$$

one takes (assuming $\|\mathbf{b}\| = 1$)

$$\mathbf{e}_m^{(k)} = V_m^{(k)} e_m(H_m^{(k)})\hat{\mathbf{e}}_1 \text{ and } b^{(k+1)} = v_{m+1}^{(k)}.$$

We will present the following three results.

Theorem 1. *Let f be a Stieltjes function, A Hermitian and positive definite, κ its condition number and assume that we take Arnoldi approximations. Then there is a constant $C > 0$ such that*

$$\|f(A)\mathbf{b} - \mathbf{f}_m^{(k)}\| \le C \cdot (\alpha_m)^k, \;\; \textit{where } \alpha_m = \frac{1}{\cosh(m \ln c)},\; c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

This means that the restarted Arnoldi method converges at a rate similar to "restarted CG" for any vector $\mathbf{b}$. Moreover, the 2-norm $\|f(A)\mathbf{b} - \mathbf{f}_m^{(k)}\|_2$ is monotonically decreasing as a function of k.

The last assertion relies on a similar result for the unrestarted method from [4].

Theorem 2. *Let f be a Stieltjes function and A be positive real, i.e. its field of values $\mathcal{F}(A)$ is contained in the right half plane. Let ν denote the distance of the origin to $\mathcal{F}(A)$ and μ its distance to $\mathcal{F}(A^{-1})$. Then, if instead of the Arnoldi approximation we take a Krylov subspace approximation inspired by the shifted GMRES approach from [5], we have*

$$\|f(A)\mathbf{b} - \mathbf{f}_m^{(k)}\|_A \leq C \cdot \beta^{mk}, \text{ where } \beta = (1 - \mu\nu)^{1/2},$$

i.e. the restarted method converges.

The transition to a Krylov subspace extraction different from the Arnoldi approximation is mandatory in Theorem 2, since we have

Theorem 3. *There are examples of Stieltjes functions f and matrices A with $\mathcal{F}(A)$ in the right half plane for which the restarted Arnoldi process does not converge for any cycle length but the maximum (full) length.*

References

[1] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, *Implementation of a restarted Krylov subspace method for the evaluation of matrix functions*, Linear Algebra Appl., 429 (2008), pp. 229–314.

[2] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, *A generalization of the steepest descent method for matrix functions*, Electron. Trans. Numer. Anal., 28 (2008), pp. 206–222.

[3] M. Eiermann and O. G. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.

[4] A. Frommer, *Monotone convergence of the Lanczos approximations to matrix functions of Hermitian matrices*, Electron. Trans. Numer. Anal., 35 (2009), pp. 118–128.

[5] A. Frommer and U. Glassner, *Restarted GMRES for shifted linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 15–26.

[6] A. Frommer, S. Güttel, and M. Schweitzer, *Efficient and stable Arnoldi restarts for matrix functions based on quadrature*, IMACM preprint, (2013).

[7] P. Henrici, *Applied and Computational Complex Analysis Vol. 2*, John Wiley & Sons, 1977.

[8] M. Ilić, I. W. Turner, and D. P. Simpson, *A restarted Lanczos approximation to functions of a symmetric matrix*, IMA J. Numer. Anal., 30 (2010), pp. 1044–1061.

[9] H. Tal-Ezer, *On restart and error estimation for Krylov approximation of $w = f(A)v$*, SIAM J. Sci. Comput., 29 (2007), pp. 2426–2441.

50 Years of Time Parallel Time Integration

Martin J. Gander

Abstract

Time parallel time integration methods have received renewed interest over the last decade because of the advent of massively parallel computers. When solving time dependent partial differential equations, the time direction is usually not used for parallelization. When parallelization in space saturates however, the time direction offers itself as a further direction for parallelization. But the time direction is special, since for evolution problems, there is a causality principle: the solution later in time is affected (it is even determined) by the solution earlier in time, and not the other way round. Algorithms trying to use the time direction for parallelization must therefore be special, and take this very different property of the time dimension into account.

I will show in this talk how time domain decomposition methods were invented, and give an overview of the existing techniques. I then explain how two classes of very successful iterative methods for steady problems, namely domain decomposition methods and multigrid methods, can be formulated for time dependent problems, and made into time parallel solvers. These solvers really do useful work at later time steps, before the solution at earlier time steps is known accurately, and thus are truly time parallel. Like in the case of steady state problems however, most of these methods are very successful for parabolic problems, but have convergence problems when applied to hyperbolic problems. We thus also discuss possible remedies for this from the more recent literature.

This talk is for people who want to quickly gain an overview of this exciting and rapidly developing area of research.

Arnoldi-Tikhonov Methods for Sparse Reconstruction

Silvia Gazzola, James Nagy and Paolo Novati

Abstract

Krylov subspace methods have been shown to be important techniques for the solution of large-scale linear discrete ill-posed problems

$$Ax + e = b\,, \quad A \in \mathbb{R}^{N\times N}, \tag{1}$$

whose available right-hand-side vector b is affected by an unknown perturbation e, and the matrix A is severely ill-conditioned, having singular values that decay to zero without a significant gap to indicate numerical rank. Arnoldi-Tikhonov (AT) methods aim at approximating the solution of the Tikhonov-regularized problem

$$\min_{x\in\mathbb{R}^N} \left\{ \|b - Ax\|_2^2 + \lambda \|L(x - x_0)\|_2^2 \right\}, \quad \lambda > 0\,, \tag{2}$$

by projecting it into Krylov subspaces of increasing dimensions generated by the Arnoldi algorithm. Arnoldi-Tikhonov methods were first proposed in [1] for the standard-form Tikhonov regularization case ($L = I$), and later generalized in many ways in order to include a generic regularization matrix $L \neq I$ (cf., [3] and the references therein). Provided that a suitable λ and L have been chosen, applying a Krylov subspace approach to solve (2) can, for many large scale applications, efficiently deliver accurate approximations of x. Note also that, if L is nonsingular, then equation (2) can be converted to standard form

$$\min_{x\in\mathbb{R}^N} \left\{ \|b - \tilde{A}\tilde{x}\|_2^2 + \lambda \|\tilde{x} - \tilde{x}_0\|_2^2 \right\}, \quad \lambda > 0\,, \tag{3}$$

where $\tilde{A} = AL^{-1}$ and $\tilde{x} = Lx$. In this formulation, we can interpret L as a right preconditioner, but the preconditioner is not chosen to accelerate convergence (e.g., by reducing the condition number of the preconditioned system); instead, L is chosen to enforce a particular regularization on the solution.

Although Tikhonov regularization can be effective, in some situations more general optimization problems of the form

$$\min_{x\in\mathbb{R}^N} \left\{ \mathcal{J}(x) + \lambda \mathcal{R}(x) \right\}, \tag{4}$$

where $\mathcal{J}(x)$ is a fit-to-data term and $\mathcal{R}(x)$ is a regularization term, can lead to higher quality reconstructions. More specifically, for many signal and image processing applications, very accurate reconstructions can be achieved by considering $\mathcal{J}(x) = \|b - Ax\|_2^2$ (exactly as in (2)) and $\mathcal{R}(x) = \|x\|_p^p$, $p \geq 1$ (in particular, if $p = 1$, sparsity is enforced into the reconstructed solution), or $\mathcal{R}(x) = \mathrm{TV}(x)$ (the total variation functional). The goal of this talk is to present two new strategies that, even if we just focus on the cases

$$\min_{x\in\mathbb{R}^N} \left\{ \|b - Ax\|_2^2 + \lambda \|x\|_1 \right\}, \tag{5}$$

$$\min_{x\in\mathbb{R}^N} \left\{ \|b - Ax\|_2^2 + \lambda \mathrm{TV}(x) \right\}, \tag{6}$$

can be potentially employed to solve problem (4) for a variety of combinations of fit-to-data and regularization terms.

The basic idea behind our methods is to incorporate an iteratively reweighted least squares (IRLS) approach into the Arnoldi-Tikhonov scheme, where the iteratively updated weighting matrices are used as regularization operators in (2). For example, in the 1-norm case, at the m-th iteration, we use the diagonal matrix $L \equiv L_m = \mathrm{diag}(1/\sqrt{|x_{m-1}|})$, where x_{m-1} is the approximate solution at the $(m-1)$-th iteration, and division, absolute value, and square-root are done element-wise.

We show how to efficiently implement this updating within the Arnoldi algorithm. Specifically, by implementing the method using the standard form (3), the "preconditioner", L_m, varies at each iteration. Therefore, in this situation, we are dealing with flexible Krylov subspace methods, and the standard Arnoldi algorithm underlying the usual formulation of the AT methods is replaced by its flexible version [4]. To the best of our knowledge, the use of flexible Krylov subspace methods for regularization purposes is novel.

In the 1-norm case, the matrices L_m are diagonal, and easy to invert. In the TV case, L_m is not easily invertible, and an inner-outer iteration scheme based on suitable restarts of the Arnoldi algorithm is proposed: basically, at the beginning of each inner cycle, the regularization matrix is updated using the solution obtained at the end of the previous inner cycle. This approach allows to heuristically enforce some additional constraints into the reconstructed solution; for instance, nonnegativity can be enforced by imposing the approximate solution at the beginning of each inner cycle to be nonnegative.

Our IRLS strategy to solve problems (5) and (6) is very close to the one described in [5], the main difference being the choice of the solution subspaces. More specifically, in [5] the normal equations associated to problem (2) are solved by a restarted CGLS method; our methods directly deal with the least squares formulation of problem (2) and, for the 1-norm case, no restarts have to be performed. Moreover, in order to choose the regularization parameter λ at each step of the Arnoldi algorithm, as well as to decide when to stop the iterations, we successfully employ an efficient strategy, based on the discrepancy principle, originally introduced in [3].

The newly proposed methods have proved to be very competitive (in terms of both computational efficiency and quality of the reconstruction) with other state-of-the-art methods commonly employed to solve problems (5) and (6). The results of some numerical experiments will be presented; we refer to [2] for further details.

References

[1] D. Calvetti, S. Morigi, L. Reichel and F. Sgallari, *Tikhonov regularization and the L-curve for large discrete ill-posed problems*. J. Comput. Appl. Math. 123, pp. 423–446, 2000.

[2] S. Gazzola and J. G. Nagy, *Generalized Arnoldi-Tikhonov method for sparse reconstruction*. Submitted, 2013.
Preprint available at `http://www.math.unipd.it/~gazzola/Sparsity.pdf` .

[3] S. Gazzola and P. Novati, *Automatic parameter setting for Arnoldi-Tikhonov methods*. J. Comput. Appl. Math. 256, pp. 180–195, 2014.

[4] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*. SIAM J. Sci. Comput. 14 (2), pp. 461–469, 1993.

[5] B. Wohlberg and P. Rodríguez, *An Iteratively Reweighted Norm algorithm for minimization of Total Variation functionals*. IEEE Signal Processing Letters 14, pp. 948–951, 2007.

High Performance Implementation of Deflated Preconditioned Conjugate Gradients with Approximate Eigenvectors

Pieter Ghysels, Wim Vanroose and Karl Meerbergen

Abstract

The method of choice for solving symmetric positive definite linear systems is typically preconditioned conjugate gradients (PCG). However, when a number of small eigenvalues are present convergence can be slow. The deflated PCG method has been proposed [1, 2] to improve convergence in this case by adding to the Krylov subspace, the space spanned by the (approximate) eigenvectors belonging to the smallest eigenvalues. In particular when the deflation space W contains the eigenvectors associated with the k smallest eigenvalues $\lambda_1, \cdots, \lambda_k$, the deflated PCG method effectively solves a linear system with condition number $\kappa = \lambda_n/\lambda_{k+1}$, with λ_n the largest eigenvalue. This is especially beneficial if there is a gap in the spectrum between λ_k and λ_{k+1}. Deflation methods are closely related to domain decomposition techniques, projection based preconditioning and coarse grid correction methods, see [3] for a discussion.

The deflated PCG method has been applied with success to a number of engineering applications like porous media flow [6] and elasticity [8]. Another important application domain is the solution of a linear system with multiple (dependent) right-hand-sides, in which case spectral information from the first solve can be used to accelerate the subsequent solves.

A direct computation of the eigenvectors for the smallest eigenvalues is prohibitively expensive, so typically approximate eigenvectors associated with a number of small eigenvalues are used as the deflation space W. There are a number of ways to obtain approximate eigenvectors, like for instance the *partial spectral factorization* [4, 5], which is based on a Chebyshev filtering iteration. In the case of multiple right-hand-sides, Ritz vectors from the first linear solve can provide approximate eigenvectors and in some cases a priori knowledge of the problem, like the geometry and the material parameters, can be used to construct crude approximations to the eigenvectors, see for instance [6] and [7].

When exact eigenvectors are available, a single oblique projection before the start of the classical PCG algorithm can be used to construct an initial guess for classical PCG such that PCG converges as if the corresponding eigenvalues where effectively removed from the spectrum. The resulting algorithm, called init-PCG [5], has the same computational cost as classical PCG, except for the single initial projection. However, when only approximate eigenvectors are available, the deflation step must be performed in every Krylov iteration, which in the case of a large deflation space can be computationally expensive. A technique like the partial spectral factorization can provide approximate eigenvectors with a user specified tolerance. In this work, we propose a heuristic to determine when (or how often) deflation is explicitly required in order to preserve the convergence of deflated PCG and save on computational cost, when only approximate eigenvectors are available.

In the classical PCG method two dot products need to be computed per iteration, requiring two global communication phases. The deflation step requires an additional global reduction. These global communication steps, typically implemented by a call to `MPI_Allreduce`, often form the bottleneck for parallel efficiency of Krylov methods. Recently, communication hiding versions of GMRES [9] and PCG [10] have been proposed that only require a single non-blocking global communication phase per iteration. In these algorithms, also called pipelined Krylov methods, the global communication can be overlapped by local work, including the matrix-vector product and possibly the application of the preconditioner. This overlapping hides both global communication

latency cost and synchronization cost. These methods show much better scalability towards large number of cores. In this work we combine deflation with the communication hiding PCG method. All global reductions, one for projection on the deflation space W and two reductions from standard PCG are combined in a single non-blocking global communication phase. We present parallel scaling results on a number of applications and a number of large parallel machines.

References

[1] Y. Saad, M. Yeung, J. Erhel and F. Guyomarc'h, *A deflated version of the conjugate gradient algorithm.* SIAM J SCI COMPUT, 21:5, 1909-1926, 2000.

[2] L. Giraud, D. Ruiz and A. Touhami, *A comparative study of iterative solvers exploiting spectral information for SPD systems.* SIAM J SCI COMPUT, 27:5, 1760-1786, 2006.

[3] J. M. Tang, R. Nabben, C. Vuik and Y. A. Erlangga, *Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods.* SIAM J SCI COMPUT, 39:3, 340-370, 2009.

[4] M. Arioli and D. Ruiz, *A Chebyshev-based two-stage iterative method as an alternative to the direct solution of linear systems.* Technical Report RAL-TR-2002-021, Rutherford Appleton Laboratory, Atlas Center, Didcot, Oxfordshire, OX11 0QX, England, 2002.

[5] L. Giraud, D. Ruiz and A. Touhami, *Krylov and polynomial iterative solvers combined with partial spectral factorization for spd linear systems.* High Performance Computing for Computational Science-VECPAR 2004. Springer Berlin Heidelberg, 2005. 637-656.

[6] C. Vuik, A. Segal, J. A. Meijerink and G. T. Wijma, *The construction of projection vectors for a Deflated ICCG method applied to problems with extreme contrasts in the coefficients.* J COMPUT PHYS., 172:2, 426-450, 2001.

[7] H. De Gersem and K. Hameyer, *A deflated iterative solver for magnetostatic finite element models with large differences in permeability.* EPJ AP, 13:1, 45-50, 2001.

[8] R. Aubry, F. Mut, S. Dey, R. Löhner, *Deflated preconditioned conjugate gradient solvers for linear elasticity.* INT J NUMER METH ENG, 88:11, 1112-1127, 2011.

[9] P. Ghysels, T. J. Ashby, K. Meerbergen and W. Vanroose, *Hiding global communication latency in the GMRES algorithm on massively parallel machines.* SIAM J SCI COMPUT, 35:1, C48-C71, 2013.

[10] P. Ghysels and W. Vanroose, *Hiding global synchronization latency in the preconditioned Conjugate Gradient algorithm.* In Press, PARALLEL COMPUT, Elsevier.

Semidefinite Programming Based Preconditioning for More Robust Near-Separable Nonnegative Matrix Factorization

Nicolas Gillis and Stephen A. Vavasis

Abstract

Nonnegative matrix factorization (NMF) under the separability assumption can be provably solved efficiently, even in the presence of noise, and has been shown to be a powerful technique in document classification and hyperspectral unmixing. This problem is referred to as near-separable NMF and requires that there exists a cone spanned by a small subset of the columns of the input nonnegative matrix approximately containing all columns. In this paper, we propose a preconditioning based on semidefinite programming making the input matrix well-conditioned. This in turn can improve significantly the performance of near-separable NMF algorithms which is illustrated on the popular successive projection algorithm (SPA). The new preconditioned SPA is provably more robust to noise, and outperforms SPA on several synthetic data sets. We also show how linear dimensionality reduction techniques and active-set methods allow us to apply the preconditioning on large-scale real-world hyperspectral images.

Keywords. nonnegative matrix factorization, semidefinite programming, preconditioning, separability, robustness to noise.

Preliminary Investigations on Recovery-Restart Strategies for Resilient Parallel Numerical Linear Algebra Solvers

E. Agullo, L. Giraud, P. Salas Medina and M. Zounon

Abstract

The advent of extreme scale machines will require the use of parallel resources at an unprecedented scale, probably leading to a high rate of hardware faults. Handling fully these faults at the computer system level may have a prohibitive cost. High performance computing applications that aim at exploiting all these resources will thus need to be resilient, *i.e.*, be able to compute a correct solution in presence of core crashes. In this work, we investigate possible remedies in the framework of numerical linear algebra problems such as the solution of linear systems or eigen-problems that are the inner most numerical kernels in many scientific and engineering applications and also ones of the most time consuming parts. More precisely, we present recovery techniques followed by restarting strategies. In the framework of Krylov subspace linear solvers the lost entries of the iterate are interpolated using the available entries on the still alive cores to define a new initial guess before restarting the Krylov method. In particular, we consider two interpolation policies that preserve key numerical properties of well-known linear solvers, namely the monotony decrease of the A-norm of the error of the conjugate gradient or the residual norm decrease of GMRES. We extend these interpolation ideas in the context of some state of the art eigensolvers where these recovery approaches are applied to reconstruct a meaningfull search space for restarting. We assess the impact of the recovery method, the fault rate and the number of processors on the robustness of the resulting numerical linear solvers. The work on the resilient linear solvers is described in [1] while the results on resilient eigensolution schemes will be reported in [2].

References

[1] E. Agullo, L. Giraud, A. Guermouche, J. Roman and M. Zounon. Towards resilient parallel linear Krylov solvers: recover-restart strategies. Research Report, Inria, RR-8324, July 2013.

[2] E. Agullo, L. Giraud, P. Salas Medina and M. Zounon. Recover-restart strategies for resilient parallel eigensolvers. Research Report, Inria, in preparation.

Extensions of the Symmetric Tridiagonal Matrix Arising from a Finite Precision Lanczos Computation

Anne Greenbaum

Abstract

Let T_J be the tridiagonal matrix that arises at some step J of a finite precision Lanczos computation for a symmetric matrix A. It was shown in [A. Greenbaum, *Behavior of Slightly Perturbed Lanczos and Conjugate-Gradient Recurrences*, Lin. Alg. Appl. 113 (1989), pp. 7–63] that T_J can be extended to a larger symmetric tridiagonal matrix,

$$T_N = \left(\begin{array}{c|ccccc} T_J & \beta_J & & & & \\ \hline \beta_J & \alpha_{J+1} & \beta_{J+1} & & & \\ & \beta_{J+1} & \ddots & \ddots & & \\ & & \ddots & \ddots & \beta_{N-1} & \\ & & & \beta_{N-1} & \alpha_N & \end{array}\right),$$

in such a way that the eigenvalues of T_N all lie in small intervals about the eigenvalues of A. If $T_N S_N = S_N \Theta_N$ is an eigendecomposition of T_N, then this implies that the tridiagonal matrix T_J would be generated at step J of exact Lanczos applied to the matrix Θ_N with initial vector equal to the first column of S_N^T. Hence whatever can be said about eigenvalue approximations generated by exact Lanczos applied to a matrix whose eigenvalues lie in small intervals about the eigenvalues of A (but where the number and exact location of these eigenvalues is not specified), can also be said about those generated by the finite precision computation for A.

The proven bound on the interval size in the 1989 paper was far from optimal. However, a procedure was given for constructing one of the infinitely many extensions to T_J whose eigenvalues would all be contained in small intervals about the eigenvalues of A. The procedure was to continue the three-term recurrence making small additional perturbations to orthogonalize new vectors against each other and against some of the Ritz vectors (the unconverged ones) from the finite precision computation. In this talk we describe a slight improvement to this construction procedure that enables us to find an extension with more tightly clustered eigenvalues. We also derive better bounds on the interval size by using the fact that perturbations to the three-term recurrence used to extend T_J lie in very special directions. Hence the eigenvalues of T_N lie even closer to those of A than the size of the perturbation terms would suggest.

Numerical Solution of Indefinite Linear Systems Arising from Interior-Point Methods

Chen Greif, Erin Moulding and Dominique Orban

Abstract

This talk focuses on the linear systems that form the core of the iterations of primal-dual interior-point methods, for solving quadratic programming problems with equality and inequality constraints. Given a symmetric and positive semidefinite Hessian matrix $H \in \mathbb{R}^{n\times n}$, vectors $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, and a Jacobian matrix $J \in \mathbb{R}^{m\times n}$, where $m \leq n$, we consider the primal-dual pair of quadratic programs (QP) in standard form

$$\min_x \; c^T x + \frac{1}{2} x^T H x \quad \text{such that} \qquad Jx = b, \quad x \geq 0, \tag{1a}$$

$$\max_{x,y,z} \; b^T y - \tfrac{1}{2} x^T H x \quad \text{such that } J^T y + z - Hx = c, \quad z \geq 0, \tag{1b}$$

where inequalities are understood elementwise, and y and z are the vectors of Lagrange multipliers associated with the equality and nonnegativity constraints of (1a), respectively. The case $H = 0$ corresponds to the linear programming (LP) problem in standard form. The distinctive feature of the class of primal-dual interior-point methods, which can be used for solving (1), is that they approximately follow a smooth path lying inside the primal-dual feasible set all the way to an optimal solution.

The primal-dual interior-point method involves a Newton iteration that requires solving linear systems of the form

$$\underbrace{\begin{bmatrix} H & J^T & -I \\ J & 0 & 0 \\ -Z & 0 & -X \end{bmatrix}}_{K} \begin{bmatrix} \Delta x \\ -\Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -c - Hx + J^T y + z \\ b - Jx \\ XZe - \tau e \end{bmatrix}. \tag{2}$$

Here τ is a barrier parameter taken to zero throughout the optimization iterations, and X and Z are diagonal matrices that become increasingly ill-conditioned as we get close to the solution of the optimization problem, i.e., as τ is reduced.

The diagonality of X and Z can be exploited to reduce the matrix size in (2) using block Gaussian elimination, and to generate either a typical block 2×2 saddle-point linear system, or reduce the system further and obtain positive definite normal equations. However, we claim that in terms of spectrum and conditioning, it may be beneficial to avoid performing such elimination steps before applying a linear solver. We use energy estimates in the spirit of [2], and obtain upper and lower bounds on the eigenvalues of the various matrices that we consider. We also consider *regularized* variants of those matrices, which allow for loosening regularity requirements.

The matrix K (and its symmetrized variant, which can be straightforwardly obtained) is nonsingular during the iterations as long as the Jacobian J is full rank. Furthermore, even at the solution itself the matrix is nonsingular under mild conditions, namely, that the solution (x, y, z) is strictly complementary, $\text{Null}(H) \cap \text{Null}(J) \cap \text{Null}(Z) = \{0\}$, and the active constraints are linearly independent. We are able to show that the eigenvalues of K are mostly bounded away from zero, with the following exceptions:

- The lower positive bound depend on the smallest eigenvalue of the Hessian and may approach zero if the Hessian is positive semidefinite;
- we are not able to obtain an effective upper negative bound which is uniformly smaller than zero.

The two above difficulties can be resolved by a technique of regularization, which proves to be very effective. We incorporate two regularization parameters, ρ and δ, and yield bounds that are uniformly bounded away from zero, with the additional benefit that there is no need to impose any requirement on the rank of the Jacobian. Our bounds lead to an asymptotic estimate of the condition number of the matrix:

$$\kappa_2(K) \lesssim 1/\min(\rho + \lambda_n, \delta),$$

where λ_n is the smallest eigenvalues of H and ρ and δ are the two regularization parameters. The dependence of the condition number on the regularization parameters is only linear, which is computationally advantageous. These results and additional observations appear in [1].

Based on the good spectral properties of the 3×3 block matrix K, we also consider preconditioning approaches for iterative solvers. Here, we show that it is possible to derive block preconditioners that are specifically tailored to the problem at hand. In particular, there exists a matrix Q such that

$$QKQ^T = \begin{bmatrix} H + \rho I & & \\ & U & \\ & & V \end{bmatrix},$$

where $H + \rho I$ is necessarily positive definite for $\rho > 0$, and U and V are negative definite. We can then flip the sign for the latter two matrices, and obtain the following useful result.

Theorem. Consider the block diagonal symmetric positive definite preconditioner

$$M = \begin{bmatrix} H + \rho I & & \\ & -U & \\ & & -V \end{bmatrix}.$$

Then, as $\tau \to 0$, the eigenvalues of the preconditioned matrix $M^{-1}K$ converge to ± 1 and $\frac{-1 \pm \sqrt{5}}{2}$. Consequently, in the absence of roundoff errors, as the solution of the optimization is approached, a minimum residual Krylov subspace solver will converge in four iterations.

The above result allows for exploiting spectral properties and strong clustering of the eigenvalues to design computationally effective iterative methods for linear systems of the form (2), for solving the optimization problem (1).

References

[1] C. Greif, E. Moulding, and D. Orban, Bounds on Eigenvalues of Matrices Arising from Interior-Point Methods, *SIAM J. Optimization*, Accepted for Publication, September 2013 (37 pages).

[2] T. Rusten and R. Winther, A Preconditioned Iterative Method for Saddlepoint Problems, *SIAM Journal on Matrix Analysis and Applications*, 13(3), 1992, pp. 887-904.

Direction Preserving Algebraic Preconditioners

Laura Grigori, Remi Lacroix, Frederic Nataf and Long Qu

Abstract

Solving large sparse linear systems of linear equations $Ax = b$ is an operation used in many industrial and academic simulations. Given the complexity of today's numerical simulations and the sizes of the linear systems that need to be solved, the design of linear solvers that have limited memory requirements and that can scale to a large number of processors is of major interest. However, most of the existing methods for solving linear systems, as domain decomposition methods with coarse space corrections or multigrid methods, are not strongly scalable, that is for a given problem size, the numerical efficiency of the preconditioner and/or its parallel performance degrades when the number of processors increases.

In this talk I will focus on a class of algebraic preconditioners that are able to preserve several directions of interest of the input matrix A. That is, given a set of vectors T which represent the directions to be preserved, the preconditioner M satisfies a right filtering property $MT = AT$. This is a property which has been exploited in different contexts, as multigrid methods [4], semiseparable matrices [9], incomplete factorizations [10, 1, 5], or nested factorization [2]. It is well known that for difficult problems with heterogeneities or multiscale physics, the iterative methods can converge very slowly, and this is often due to the presence of several low frequency modes. By preserving the directions corresponding to these low frequency modes in the preconditioner, their effect on the convergence is alleviated and a much faster convergence is often observed.

In our work we have developed two different direction preserving algebraic preconditioners. The first one, BFD [5], is based on a block decomposition of the input matrix. The second one, NFF, is based on a recursive decomposition that requires first to permute the input matrix, which can have an arbitrary sparsity structure, into a matrix with a nested block arrow structure. This recursive factorization is a key feature in allowing NFF to have limited memory requirements and also to be very well suited for hierarchical parallel machines. The construction of both preconditioners involve the approximation of Schur complements of the form $LD^{-1}U$ that appear at each step of the recursive decomposition. Different preconditioners can be obtained by using different approximations, and more details can be found in [6]. The specific approximation used in NFF that allows to satisfy the right filtering property was introduced in [8] and is similar to the approach used in BFD [7, 5], but adapted to a recursive computation.

Our talk will focus on introducing the two preconditioners and discussing their numerical properties. In particular we will show that when the input matrix is symmetric and positive definite, the preconditioner also remains symmetric and positive definite. We will also discuss the convergence of NFF on a set of matrices arising from the discretization of a boundary value problem with highly heterogeneous coefficients on three-dimensional grids. Our results show that on a $400 \times 400 \times 400$ regular grid, the number of iterations with NFF increases slightly while increasing the number of subdomains up to 2048. In terms of runtime performance on Curie, a Bullx system formed by nodes of two eight-core Intel Sandy Bridge processors, our preconditioners scale well up to 1024 cores and it is 2.6 times faster than the domain decomposition preconditioner Restricted Additive Schwarz (RAS) as implemented in PETSc [3].

The choice of the filtering vectors plays an important role in direction preserving preconditioners. There are problems for which we have prior knowledge of the near kernel of the input matrix, and this is indeed the case for some of the problems that we will be presenting in the talk. They can

also be approximated by using techniques similar to the ones used in deflation, and this aspect will be also addressed in our talk.

References

[1] Y. Achdou and F. Nataf. An iterated tangential filtering decomposition. *Numer. Linear Algebra Appl.*, 10(5-6):511–539, 2003. Preconditioning, 2001 (Tahoe City, CA).

[2] J. Appleyard and I. Cheshire. Nested factorization. In *Seventh SPE Symposium on Reservoir Simulation*, pages 315–324, 1983. paper number 12264.

[3] S. Balay, J. Brown, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. PETSc Web page, 2012. http://www.mcs.anl.gov/petsc.

[4] A. Brandt, J. Brannick, K. Kahl, and I. Livshits. Bootstrap AMG. *SIAM J. Sci. Comput.*, 33(2):612–632, 2011.

[5] R. Fezzani, L. Grigori, F. Nataf, and K. Wang. Block Filtering Decomposition. In *Numerical Linear Algebra with Applications (NLAA)*, submitted 2012.

[6] L. Grigori, P. Kumar, F. Nataf, and K. Wang. A class of multilevel parallel preconditioning strategies. Research Report RR-7410, INRIA, Oct. 2010.

[7] L. Grigori and F. Nataf. Generalized Filtering Decomposition. Research Report RR-7569, INRIA, Mar. 2011.

[8] L. Grigori, F. Nataf, and P. Kumar. Multipurpose Calculation Computing Device, 2010. FR Patent WO/2012/035272 filed September 17, 2010, International Application No. PCT/FR2011/052128.

[9] M. Gu, X. S. Li, and P. S. Vassilevski. Direction-preserving and Schur-monotonic semiseparable approximations of symmetric positive definite matrices. *SIAM J. Matrix Anal. Appl.*, 31(5):2650–2664, 2010.

[10] C. Wagner. Tangential frequency filtering decompositions for symmetric matrices. *Numer. Math.*, 78(1):119–142, 1997.

Rapid Convergence for Finite Rank Approximations of Infinite-dimensional Lyapunov Equations

Luka Grubišić and Daniel Kressner

Abstract

We analyze the convergence properties of an explicitly constructed sequence of low rank approximations to the solution of an infinite dimensional operator Lyapunov equation. As an application of our abstract theory we consider approximations of linear control system in the context of model reduction by balanced truncation.

In particular we are interested in systems governed by the heat equation with both distributed as well as boundary control. For such systems in a recent study [7] the authors have presented an ADI-based algorithm in infinite dimensional setting which was used to construct low rank approximations of the solutions of a Lyapunov equation. A particular emphasis in [7] was on allowing Lyapunov equations with all unbounded coefficients.

Let now A and B be unbounded operators. The formal expression $AX + XA' = -BB'$, where A' and B' are appropriate operator duals, is called an abstract Lyapunov equation. Following [5] we analyze the approximation properties of solutions of abstract Lyapunov equations posed in the setting of a scale of Hilbert spaces associated to an unbounded diagonalizable operator which satisfies the Kato's square root theorem [1]. We call an (unbounded) operator A diagonalizable if there exists a bounded operator Q, with a bounded inverse, such that the (unbounded) operator $Q^{-1}AQ$ is a normal operator with a compact resolvent. In this setting we assume that $-A$ satisfies the requirements of [1].

Let further $\sup(\mathsf{Re}(\mathrm{Spec}(A))) < 0$ and let $(-A)^{-1/2}B$ be bounded, then the abstract Lyapunov equation has a unique minimal positive and bounded solution operator X. We further assume that the input space of the linear system is finite dimensional. This amounts to an assumption that B' has finite rank relatively to this Hilbert space structure.

Under these assumptions an abstract Lyapunov equation can be interpreted as a weak operator equation from [4]. Subsequently, we combine explicit representation formulas for the solutions of weak Lyapunov equations with tensor approximation techniques from [3, 6] to obtain a sequence of rapidly converging low rank approximations to X. We also obtain an explicit estimate for the exponential decay of the singular values $\sigma_i(X)$, $i \in \mathbb{N}$ (see also [8]).

To this end, we construct a degenerate—of rank $(2k+1) \times \mathrm{Ran}(B)$—approximation X_{2k} to the operator X and give conditions under which an estimate

$$\sqrt{\sum_{i=2k+2}^{\infty} \sigma_i^2(X)} \leq \|X - X_{2k}\|_{HS} \leq O(\exp^{-\pi\sqrt{2k}}) \,, \tag{1}$$

holds. Here $\|X\|_{HS} = \sqrt{\mathrm{tr}\,(X^*X)}$ denotes the Hilbert-Schmidt norm of X and we note that our technique allows us to give the constant in the $O(\cdot)$ notation explicitly in terms of the weighted norms of A, B and Q and an estimate on the spectrum of A.

In the case of a (more strongly) unbounded control operator B, e.g. an operator only bounded in a weighted Hilbert space, we obtain the same type of convergence estimates in an associated weighted norm on (the subspace of) the space of compact operators.

Our numerical experiments are designed to test the sharpness of the estimate (1). We discretize the infinite dimensional Lyapunov equation by a Galerkin projection using both finite element as well as spectral element methods. We point out that the fact that $-A$ satisfies the Kato's square root theorem from [1] was essential in justifying the use of finite elements with only first order Sobolev regularity, eg. Lagrange P1 elements for the heat equation, when projecting the Lyapunov equation. We then use the extended Arnoldi algorithm, see [2, 9], to efficiently compute low rank approximations to the solution operator X.

Based on our convergence estimates we also discuss ramifications of this analysis for the design of adaptive finite element methods including the analysis of the influence of linear algebra approximations on the overall process.

References

[1] A. Axelsson, S. Keith, and A. McIntosh. The Kato square root problem for mixed boundary value problems. *J. London Math. Soc. (2)*, 74(1):113–130, 2006.

[2] V. Druskin, L. Knizhnerman, and V. Simoncini. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. *SIAM J. Numer. Anal.*, 49(5):1875–1898, 2011.

[3] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

[4] L. Grubišić and K. Veselić. On weakly formulated Sylvester equations and applications. *Integral Equations Operator Theory*, 58(2):175–204, 2007.

[5] B. Jacob and J. R. Partington. On controllability of diagonal systems with one-dimensional input space. *Systems Control Lett.*, 55(4):321–328, 2006.

[6] D. Kressner and C. Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.*, 31(4):1688–1714, 2010.

[7] M. Opmeer, T. Reis, and W. Wollner. Finite-rank ADI iteration for operator Lyapunov equations. *SIAM J. Control Optim.*, 51(5):4084–4117, 2013.

[8] M. R. Opmeer. Decay of hankel singular values of analytic control systems. *Systems & Control Letters*, 59(10):635 – 638, 2010.

[9] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput.*, 29(3):1268–1288, 2007.

Perfectly Matched Layers via the Iterated Rational Krylov Algorithm

Vladimir Druskin, Stefan Güttel and Leonid Knizhnerman

Abstract

An important task in science and engineering is the numerical solution of a partial differential equation (PDE) on an unbounded domain. Complex coordinate stretching is a powerful idea for constructing computational grids that, in some sense, mimic unbounded domains, with probably the most popular approach being known as *perfectly matched layer* (PML, see the influential work of Berenger [2]). In this talk we will present a new approach for constructing such layers. It is based on ideas of the eminent mathematicians Y. I. Zolotarev (1847–1878), T. J. Stieltjes (1856–1894), and M. G. Krein (1907–1989), and modern linear algebra techniques.

To outline the main idea, let us consider the following two-point boundary value problem

$$\frac{\partial^2}{\partial x^2}\mathbf{u} = \mathbf{A}\mathbf{u}, \quad \frac{\partial}{\partial x}\mathbf{u}\Big|_{x=0} = -\mathbf{b}, \quad \mathbf{u}\Big|_{x=+\infty} = \mathbf{0}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{N\times N}$ is nonsingular and $\{\mathbf{b}, \mathbf{u}(x)\} \subset \mathbb{C}^N$. If $\mathbf{A}$ is a discretization of a differential operator on some spatial domain $\Omega \subseteq \mathbb{R}^\ell$, then (1) is a semidiscretization of an $(\ell+1)$-dimensional PDE on $[0,+\infty)\times\Omega$. Assuming that problem (1) is well posed, its exact solution can be given in terms of matrix functions as $\mathbf{u}(x) = \exp(-x\mathbf{A}^{1/2})\mathbf{A}^{-1/2}\mathbf{b}$. In particular, at $x = 0$ the solution is given as

$$\mathbf{u}(0) = F(\mathbf{A})\mathbf{b}, \quad F(z) = z^{-1/2}. \tag{2}$$

The relation (2) allows for the exact conversion of Neumann data $-\mathbf{b}$ at the boundary $x = 0$ into the Dirichlet data $\mathbf{u}(0)$, without the need for solving (1) on its unbounded domain.

A difficult, but practically important case is obtained when $\mathbf{A}$ in (2) is a Hermitian indefinite matrix. For example, when solving wave scattering problems one typically deals with a discretization of the negative shifted Laplacian $-\Delta - k^2$ on $\Omega \subset \mathbb{R}^\ell$, in which case problem (1) is a semidiscretization of the indefinite Helmholtz equation on $[0,+\infty)\times\Omega$. It is thus reasonable to assume that

$$\mathbf{A} = \mathbf{L} - k^2\mathbf{I},$$

where $\mathbf{L} \in \mathbb{C}^{N\times N}$ is Hermitian positive definite, $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix, and $k^2 > 0$ is not in the spectrum of $\mathbf{L}$. In a recent work [3] we have presented the construction of a near-optimal rational approximant $R_n(z)$ to the function $F(z)$ defined in (2) such that $R_n(\mathbf{A}) \approx F(\mathbf{A})$, with the branch cut of $F(z)$ modified appropriately. Let us assume that the real eigenvalues of $\mathbf{A}$ are ordered such that

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_i < 0 < \lambda_{i+1} \le \cdots \le \lambda_N.$$

Our construction in [3] required estimates for the eigenvalues $\lambda_1, \lambda_i, \lambda_{i+1}, \lambda_N$, i.e., we needed to know the endpoints of the negative and positive spectral subintervals $\Sigma_1 = [\lambda_1, \lambda_i]$ and $\Sigma_2 = [\lambda_{i+1}, \lambda_N]$ of $\mathbf{A}$. The rational approximant $R_n(z)$ for $F(z)$ was then constructed by combining two relative best rational approximants related to Σ_1 and Σ_2, and it minimizes the relative error $|1 - R_n(z)/F(z)|$ uniformly on $\Sigma_1\cup\Sigma_2$. While the outer eigenvalues λ_1, λ_N are easily estimated, the inner eigenvalues λ_i, λ_{i+1} are typically more difficult to obtain.

In a more recent work we describe a new technique for constructing a different rational approximant $R_n(z)$ to the function $F(z)$ defined in (2). We will make use of the fact that any type $(n-1, n)$

rational interpolant $R_n(z)$ for $F(z)$ can be represented as

$$R_n(z) = \frac{P_{n-1}(z)}{Q_n(z)} = F(z) \cdot \frac{H_m(-\sqrt{z}) - H_m(\sqrt{z})}{H_m(-\sqrt{z}) + H_m(\sqrt{z})},$$

where H_m is a monic polynomial of degree $m = 2n$. The m interpolation conditions are satisfied at points z_j for which $H_m(\sigma_j) = 0$, $z_j = \sigma_j^2$ $(j = 1, \ldots, m)$. The denominator $Q_n(z)$ is determined from the even part of $H_m(z)$, and the numerator $P_{n-1}(z)$ from the odd part. In order to compute optimal interpolation nodes z_j we run the *iterated rational Krylov algorithm* for optimal $\mathcal{H}_2$ model reduction [4] with the matrix $\mathbf{S} = \sqrt{\mathbf{A}}$. Our algorithm can be sketched as follows:

1. Start with some (distinct) initial shifts $\sigma_1, \ldots, \sigma_m$.
2. Compute an orthonormal basis $\mathbf{V}_m \in \mathbb{C}^{N \times m}$ of the m-dimensional rational Krylov space
$$\mathcal{Q}_m(\mathbf{S}, \mathbf{b}) = \text{span}\{(\mathbf{S} - \sigma_1 \mathbf{I})^{-1}\mathbf{b}, \ldots, (\mathbf{S} - \sigma_m \mathbf{I})^{-1}\mathbf{b}\}.$$
3. Compute the projection $\mathbf{S}_m = \mathbf{V}_m^* \mathbf{S} \mathbf{V}_m \in \mathbb{C}^{m \times m}$ and set $\{\sigma_1, \ldots, \sigma_m\} := -\Lambda(\mathbf{S}_m)$.
4. Repeat from step 2. until convergence.
5. The required interpolation nodes are $z_j = \sigma_j^2$ $(j = 1, \ldots, m)$.

In this algorithm only the actions of shifted and inverted versions of $\mathbf{S}$ onto vectors are required, which themselves can be computed efficiently by rational Krylov techniques [5]. The resulting interpolant $R_n(z)$ is parameter-free (except for its degree n, of course), so its construction does not require estimates for the spectral subintervals of $\mathbf{A}$. Moreover, using a result from [1], we can characterize how the $\mathcal{H}_2$ optimality of the shifts σ_j for $\mathbf{S}$ translates into an optimality condition for the interpolant $R_n(z)$.

The main features of this new interpolant compared to our previous work [3] are that its construction is entirely based on linear algebra techniques, and that it possesses "spectral awareness", i.e., it is optimized for the discrete set of eigenvalues of $\mathbf{A}$ and weights given by the vector $\mathbf{b}$. This discrete optimality can result in significant reduction of n, the degree of the interpolant, for achieving a given error tolerance. In particular the last advantage is crucial for the construction of perfectly matched layers from $R_n(z)$ with fewest possible grid points n, thereby reducing significantly the overall size of the discretization matrices for 3D Helmholtz problems on unbounded domains.

References

[1] P. Benner and T. Breiten, *On optimality of interpolation-based low rank approximations of large scale matrix equations,* Preprint, 2012.

[2] J. P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comp. Phys., 114 (1994), pp. 185–200.

[3] V. Druskin, S. Güttel, and L. Knizhnerman, *Near-optimal perfectly matched layers for indefinite Helmholtz problems*, The University of Manchester, MIMS Eprint, 24 pages, submitted 2013.

[4] S. Gugercin, *An iterative rational Krylov algorithm (IRKA) for optimal $\mathcal{H}_2$ model reduction*, Householder Symposium XVI, Seven Springs Mountain Resort, PA, USA, 2005.

[5] S. Güttel and L. Knizhnerman, *A black-box rational Arnoldi variant for Cauchy-Stieltjes matrix functions,* BIT Numer. Math., 53 (2013), pp. 595–616.

The Sylvester Equation and Interpolatory Model Reduction of Linear/Bilinear Dynamical Systems

Garret Flagg and Serkan Gugercin

Abstract

Consider a single-input single-output (SISO) linear dynamical system in the state-space form

$$\dot{x}(t) = Ax(t) + bu(t), \quad y(t) = c^T x(t) \quad \Longleftrightarrow \quad G(s) = c^T(sI - A)^{-1}b \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}$ and $y \in \mathbb{R}$ are, respectively, the states, input and output of the underlying system; $A \in \mathbb{R}^{n\times n}$, and $b, c \in \mathbb{R}^n$. In (1), $G(s)$ is called the transfer function. We will assume that the dynamical system is asymptotically stable; i.e., the eigenvalues of A have negative real parts. For the cases where the system order, i.e., the state dimension n, is very large, one would like to approximate the full-order model (1) with a reduced model of similar form. Thus, the goal is to construct a reduced model

$$\dot{x}_r(t) = A_r x_r(t) + b_r u(t), \quad y_r(t) = c_r^T x_r(t) \quad \Longleftrightarrow \quad G_r(s) = c_r^T(sI_r - A_r)^{-1}b_r \tag{2}$$

where $x_r \in \mathbb{R}^r$ with $r < n$, $A_r \in \mathbb{R}^{r\times r}$, and $b_r, c_r \in \mathbb{R}^r$ such that the reduced transfer function $G_r(s)$ is a good approximation to $G(s)$ in an appropriate norm. A commonly used method for obtaining $G_r(s)$ is rational interpolation. Given the interpolation points $\{\sigma_i\}_{i=1}^r \in \mathbb{C}$, the goal is to construct $G_r(s)$ such that

$$G_r(\sigma_i) = G(\sigma_i) \qquad \text{and} \qquad G_r'(\sigma_i) = G'(\sigma_i) \qquad \text{for} \;\; i = 1, 2, \ldots, r, \tag{3}$$

where $'$ denotes differentiation with respect to s. Here, we focus on the simple Hermite interpolation; however the discussion can be generalized to matching the higher-order derivatives as well; see [3] for a recent survey paper on model reduction by interpolation. We will assume that the interpolation points lie in the right-half plane.

The Sylvester equation plays a fundamental role in this framework. Given the interpolation points $\{\sigma_i\}_{i=1}^r$, let $V \in \mathbb{C}^{n\times r}$ and $W \in \mathbb{C}^{n\times r}$ solve the following two Sylvester equations:

$$AV - V\Sigma = b\,e^T \quad \text{and} \quad A^T W - W\Sigma = c\,e^T \tag{4}$$

where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{C}^{r\times r}$ and $e \in \mathbb{R}^r$ is the vector of ones. Then, the reduced model $G_r(s) = c_r^T(sI_r - A_r)^{-1}b_r$ given by $A_r = (W_r^T V_r)^{-1}W_r^T A V_r$, $b_r = (W_r^T V_r)^{-1}W_r^T b$, and $c_r = V_r^T c$ satisfies the interpolation conditions given in (3), see, e.g., [7]. The Sylvester equation has proved important in optimal model reduction as well. For example, if the goal is to find a reduced model to minimize the H_2 error norm, defined as $\|G - G_r\|_{H_2} = \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\imath\omega) - G_r(\imath\omega)|^2 d\omega}$, one can solve an iteratively updated sequence of Sylvester equations of the form (4) to construct a locally optimal H_2 approximation ([2]). This framework is equivalent to the Iterative Rational Krylov Algorithm [6] framework for optimal H_2 model reduction as discussed in [2, 6].

Now, consider a SISO bilinear system: $\dot{x}(t) = Ax(t) + Nx(t)u(t) + bu(t), \quad y(t) = c^T x(t)$, where $A, N \in \mathbb{R}^{n\times n}$, and $b, c \in \mathbb{R}^n$. The bilinearity is due to the term $Nx(t)u(t)$ in the state equation. The main question we want to answer in this talk is how to extend the connection between the Sylvester equation and interpolation to the bilinear setting where the reduced bilinear model will have the state-space form: $\dot{x}_r(t) = A_r x_r(t) + N_r x_r(t)u(t) + b_r u(t), \quad y_r(t) = c_r^T x_r(t)$,

with $A, N \in \mathbb{R}^{r\times r}$, and $b, c \in \mathbb{R}^r$. The Volterra series representation of bilinear systems fills a gap left by the absence of transfer functions. The output $y(t)$ can be expressed as $y(t) = \sum_{k=1}^{\infty} \int_0^t \int_0^{t_1} \cdots \int_0^{t_k-1} \widetilde{g}(t_1, \ldots, t_k) u(t-t_1 \cdots - t_k) \cdots u(t-t_k) dt_k \cdots dt_1$, assuming the conditions for the Volterra series to converge. By applying a multivariate Laplace transform to $\widetilde{g}(t_1, \ldots, t_k)$, one obtains the k^{th} subsystem transfer function $\widetilde{G}_k(s_1, s_2, \ldots, s_{k-1}, s_k)$. Interpolation methods for bilinear systems have so far mainly focused on interpolating some of the leading k^{th} transfer functions (see, e.g., [1, 5]). In this talk, we will show how to interpolate the underlying Volterra series.

Given two sets of interpolation points $\{\sigma_i\}_{i=1}^r \in \mathbb{C}$ and $\{\mu_i\}_{i=1}^r \in \mathbb{C}$, together with two matrices $U, S \in \mathbb{R}^{r\times r}$, for which convergence of certain infinite series is assumed, let $V \in \mathbb{C}^{n\times r}$ and $W \in \mathbb{C}^{n\times r}$ solve the following two bilinear Sylvester equations

$$V_r\Sigma - AV_r - NV_rU^T = be^T \quad \text{and} \quad W_rM - A^TW_r - N^TV_rS^T = ce^T, \tag{5}$$

where $M = \text{diag}(\mu_1, \ldots, \mu_r) \in \mathbb{C}^{r\times r}$. Then we will show that the reduced bilinear model constructed as $A_r = (W_r^TV_r)^{-1}W_r^TAV_r$, $N_r = (W_r^TV_r)^{-1}W_r^TNV_r$, $b_r = (W_r^TV_r)^{-1}W_r^Tb$, and $c_r = V_r^Tc$ *interpolates the full Volterra series*

$$\sum_{k=1}^{\infty} \sum_{l_1}^{r} \sum_{l_2}^{r} \cdots \sum_{l_{k-1}}^{r} \eta_{l_1,l_2,\ldots,l_{k-1},j} \widetilde{G}_k(\sigma_{l_1}, \sigma_{l_2}, \ldots, \sigma_j)$$

for every $\{\sigma_j\}_{j=1}^r$ where the weights $\eta_{l_1,l_2,\ldots,l_{k-1},j}$ depends on U. Similar result holds for interpolation at μ_j. This extends the connection between rational interpolation and the Sylvester equation to bilinear systems by focusing on interpolating the underlying Volterra series. Benner and Breiten [4] has recently extended the optimal H_2 model reduction to bilinear systems where an iteratively updated sequence of (bilinear) Sylvester equations of the form (5) are solved. With the new connection between the Sylvester equation and the Volterra series interpolation, we will also show that [4] achieves interpolation in the Volterra series at carefully selected interpolation points.

References

[1] Z. Bai and D. Skoogh. *A projection method for model reduction of bilinear dynamical systems.* Linear Algebra and its Appl., Vol. 415, No: 2-3. pp. 406–425, 2006.

[2] P. Benner, M. Köhler, and J. Saak. *Sparse-Dense Sylvester Equations in H_2 Model Order Reduction.* Max Planck Institute Magdeburg Preprints MPIMD/11-11, 2011.

[3] A.C. Antoulas, C.A. Beattie and S. Gugercin. *Interpolatory model reduction of large-scale dynamical systems.* Efficient Modeling and Control of Large-Scale Systems, J. Mohammadpour and K. Grigoriadis editors, Springer-Verlag, 2010.

[4] P. Benner and T. Breiten. *Interpolation-Based H_2-Model Reduction of Bilinear Control Systems.* SIAM J. on Matrix Anal. and Appl., Vol. 30; Issue: 3, pp. 859-885, 2012.

[5] T. Breiten and T. Damm. *Krylov subspace methods for model order reduction of bilinear control systems*, Systems & Control Letters, Vol. 59, Issue 8, pp. 443–450, 2010.

[6] S. Gugercin, A.C. Antoulas and C.A. Beattie, H_2 model reduction for large-scale linear dynamical systems. SIAM J. on Matrix Anal. and Appl., Vol. 30, Issue: 2, pp. 609-638, 2008.

[7] K. Gallivan, A. Vandendorpe, P. Van Dooren. *Sylvester equations and projection-based model reduction*, J. of Computational and Applied Mathematics, Vol. 162. Issue 1, pp. 213-229, 2004.

Performance Enhancement of Doubling Algorithms for a Class of Complex Nonsymmetric Algebraic Riccati Equations

Chun-Hua Guo, Changli Liu and Jungong Xue

Abstract

We consider the nonsymmetric algebraic Riccati equation (NARE)

$$XCX - XD - AX + B = 0, \tag{1}$$

where A, B, C, D are complex matrices of sizes $m \times m, m \times n, n \times m, n \times n$, respectively. Associated with the NARE (1) is the matrix

$$Q = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix}. \tag{2}$$

The NARE (1) is said to be in class H^* if the comparison matrix $\widehat{Q}$ of Q, defined by

$$[\widehat{Q}]_{ij} = \begin{cases} \mathrm{Re}([Q]_{ii}), & i = j, \\ -|[Q]_{ij}|, & i \neq j, \end{cases}$$

is a nonsingular M-matrix. The class H^* is an extension of the class H^+ studied earlier in [2], where the diagonal entries of Q are required to be real and positive. The NARE in class H^* arises in the study of Markov modulated fluid flows; see [5] and the references therein.

The study of the NARE in class H^+ or H^* is through comparison with a NARE (1) for which the matrix Q in (2) is a nonsingular M-matrix. The later is said to be in class M or called an M-matrix algebraic Riccati equation, which has been studied extensively (see [1, 3], for example). Any NARE in class M has a minimal nonnegative solution. In the study of the NARE (1) in class H^*, we may assume without loss of generality that $\widehat{Q}\mathbf{1} > 0$, where $\mathbf{1}$ is the vector of ones. We have the following result of [5], which is a useful generalization of [2, Theorem 8].

Theorem 1. *Suppose the NARE* (1) *is in class* H^* *and* $\widehat{Q}\mathbf{1} > 0$. *Let* $\widetilde{\Phi}$ *be the minimal nonnegative solution of the NARE*

$$X\widetilde{C}X - X\widetilde{D} - \widetilde{A}X + \widetilde{B} = 0,$$

where

$$\widetilde{Q} = \begin{bmatrix} \widetilde{D} & -\widetilde{C} \\ -\widetilde{B} & \widetilde{A} \end{bmatrix},$$

partitioned as for Q *in* (2), *is a nonsingular* M*-matrix satisfying* $\widetilde{Q} \leq \widehat{Q}$ *and* $\widetilde{Q}\mathbf{1} > 0$. *Then the NARE* (1) *has a unique solution* Φ *such that* $|\Phi| \leq \widetilde{\Phi}$. *Similarly, the dual equation of* (1)

$$YBY - YA - DY + C = 0$$

has a unique solution Ψ *such that* $|\Psi| \leq \widetilde{\Psi}$, *where* $\widetilde{\Psi}$ *is the minimal nonnegative solution of the NARE*

$$Y\widetilde{B}Y - Y\widetilde{A} - \widetilde{D}Y + \widetilde{C} = 0.$$

In [5] it is shown that the special solutions Φ and Ψ in Theorem 1 are the solutions required in applications and that these two solutions can be found simultaneously by existing doubling algorithms [4, 6] if the parameters in the doubling algorithms are chosen properly. In this talk,

we show that the performance of the doubling algorithms can often be improved significantly if a proper preprocessing procedure is used on the given Riccati equation, at a negligible cost. We also propose new strategies for choosing parameters for doubling algorithms. For some difficult cases, these strategies can provide significant further improvement after using the preprocessing procedure.

References

[1] C.-H. Guo, *Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 225–242.

[2] C.-H. Guo, *A new class of nonsymmetric algebraic Riccati equations*, Linear Algebra Appl., 426 (2007), pp. 636–649.

[3] C.-H. Guo and N. J. Higham, *Iterative solution of a nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 396–412.

[4] X.-X. Guo, W.-W. Lin, and S.-F. Xu, *A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation*, Numer. Math., 103 (2006), pp. 393–412.

[5] C. Liu and J. Xue, *Complex nonsymmetric algebraic Riccati equations arising in Markov modulated fluid flows*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 569–596.

[6] W.-G. Wang, W.-C. Wang, and R.-C. Li, *Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 170–194.

Is There a Market for Modified Moments?

Martin H. Gutknecht

Abstract

What the engineers call 'partial realization' is known to mathematicians and physicists as (matrix) Padé approximation at ∞ of the transfer function $H(s) = C(sI - A)^{-1}B$ of a linear time-invariant system. The approach is also known as moment (or Markov parameter) matching. Traditionally, this matching has been achieved by solving a Hankel or block Hankel system of linear equations, typically achieved by fast recursive algorithms. But already in the mid-50s Rutishauser became aware of the ill-conditioning of this moment matching, which hampered his qd algorithm. So, for computing the continued fraction of a scalar transfer function H, he suggested to apply the Lanczos algorithm, and for computing the poles of H he would subsequently apply the progressive form of his qd algorithm, which is the same as applying his LR algorithm to the tridiagonal matrix of the Lanczos recurrence coefficients. So, in the scalar case, the computation of Padé approximations at ∞ was introduced nearly 40 years before Padé-via-Lanczos (PVL) became widely used in the control community following the publications of Feldmann and Freund (1994, 1995). In the 1970s and 1980s such applications of the Lanczos algorithm were also much promoted by G.H. Golub and W.B. Gragg. However, all these algorithms can break down if A is not Hpd.

Another efficient but unstable alternative to solving a Hankel system for moment matching had been known long before: the Chebyshev algorithm (1859), which, in fact, can also be viewed as a fast Hankel solver providing the recursions of the corresponding orthogonal polynomials. In the 1960s Gautschi linked the instability of the Chebyshev algorithm to the ill-conditioning of the moment matching, and he also showed that the so-called modified Chebyshev algorithm of Sack and Donovan (1972) and Wheeler (1974) may behave much better. However, also the modified Chebyshev algorithm can break down in the same way the nonsymmetric Lanczos algorithm can break down, because it produces the same continued fraction and the same Padé approximants.

In 1990, Golub and Gutknecht came up with a version of this modified Chebyshev algorithm that could overcome breakdowns in exact arithmetic. However, unlike the look-ahead Lanczos algorithm this 'reliable' or 'non-generic' modified Chebyshev algorithm does not remain stable in the case of a near-breakdowns in finite-precision arithmetic, and its extension to a look-ahead algorithm is not at all straightforward. The first aim of our renewed interest in this area was to fill this long-standing gap. Achieving it turned out to be a bit tricky, but simpler than expected. The resulting look-ahead modified Chebyshev algorithm generates (in the scalar, SISO case) the same sequence of block tridiagonal upper Hessenberg matrices as the look-ahead Lanczos algorithm. These matrices are Petrov–Galerkin projections of A.

Other challenges remain: what about the MIMO case? What about rational interpolation instead of Padé approximation at ∞? Moreover: what about applications? Golub's main intention, realized in the PhD thesis of Mark Kent (1989), was to use the modified Chebyshev algorithm for first approximating the extremal eigenvalues of an spd matrix A and then to use this information for determining the parameters of the Chebyshev iteration, a classical Krylov solver for matrices with positive real spectrum. Today, computing approximate eigenvalues with the modified Chebyshev algorithm may still be competitive, in particular since the algorithm is naturally communication-avoiding and not restricted to the Hermitian case. Can it be made more stable than Lanczos? As we indicated, this modified Chebyshev algorithm can also be applied for model reduction by partial realization. Yet other applications may get into the focus. For example, using the resulting spectral information for the augmentation and deflation of Krylov subspaces.

Fast Nonstationary Preconditioned Iterative Methods for Image Deblurring

Marco Donatelli and Martin Hanke

Abstract

For large-scale discrete linear ill-posed problems

$$Tx = y\,, \tag{1}$$

as they arise, for example, in image deblurring applications, iterative regularization methods provide a welcome alternative to, say, Tikhonov regularization, because iterative methods are comparatively cheap to implement and regularize the problem "on the fly". The regularization parameter of these methods is the stopping index, that means, regularization is achieved by early stopping of the iteration. This avoids the cumbersome trial and error process known from Tikhonov regularization, where large linear systems have to be solved for a set of conceivable regularization parameters before the final output is determined.

While the basic Landweber iteration (i.e., the fixed point iteration for the normal equation system) is far too slow to be useful, the conjugate gradient iteration (CGLS) is a much more valuable option for this purpose. But even CGLS may take a few tens of iterations to achieve good accuracy, and in the end this may tip the scales towards cheaper Fourier based methods with inferior quality.

Because of that it has been suggested in the 90's to utilize Fourier techniques to design preconditioners to speed up the iteration, cf., e.g., [3, 4, 5]. When doing so, care has to be taken that the preconditioner does not spoil the benefits of the iterative process: if the preconditioner comes without any regularization then its action will inevitably introduce disastrous noise propagation into the iterates; on the other hand, excessive regularization or filtering prevents the reconstruction of relevant signal components.

Another issue of the implementation of the CGLS iteration (be it preconditioned or not) is the proper choice of the stopping index (i.e., the regularization parameter). This choice has to be based on metadata, such as the norm of the residual and good a priori knowledge about the noise level, or on human interaction and excellent pertinent experience. A bad choice of the stopping index will have the same effect as a careless implementation of the preconditioner described above.

Because of these difficulties preconditioned iterative regularization methods for image deblurring problems still suffer under a lack of robustness. In this work, cf. [1], we therefore suggest a new iterative method that avoids the conjugate gradient iteration, by utilizing a very powerful nonstationary (Fourier based) preconditioning scheme for the Landweber iteration instead. The method shares similarities to nonstationary iterated Tikhonov regularization without the corresponding expensive system solves.

To be specific the algorithm has the following form. Let C be a suitable approximation (see below) of the system matrix T, and x_0 be an initial guess of the exact solution. Then, for $n = 0, 1, 2, \ldots$, compute

$$x_{n+1} = x_n + h_n\,, \tag{2}$$

where h_n is given by

$$h_n = C^*(CC^* + \alpha_n I)^{-1}(y - Tx_n)\,. \tag{3}$$

Here, $\{\alpha_n\}$ is a sequence of positive regularization parameters, to be determined adaptively by stipulating

$$\|r_n - Ch_n\| \;=\; q_n\|r_n\|\,, \tag{4}$$

where $r_n = y - Tx_n$ is the residual, and q_n is taken to be a given parameter ($q_n = 0.8$, say) in the beginning of the iteration, and increasing towards one in the final steps of the iteration. Take note that the approximation C comes without any regularization; instead, the corresponding regularization is incorporated explicitly via the parameters α_n. Also note that the need for efficient means for solving (3), (4) provides severe restrictions on suitable choices for C; as mentioned before, we use Fourier based approximations of T in our examples, and then the dominating work load of a single step of the iteration is the computation of the residual r_n.

The overall process (2)–(4) has the spirit of an inexact Newton scheme – despite the fact that the underlying system (1) is linear. In fact, our theoretical analysis of this iteration employs techniques that have been developed for an analysis of a regularizing Levenberg-Marquardt iteration for nonlinear ill-posed problems [2]. The key ingredient of this analysis is a closeness assumption on C, namely that

$$\|(C - T)z\| \leq \rho \|Tz\| \tag{5}$$

for some $0 < \rho < 1/2$ and all z.

Our theoretical results include a linear rate of convergence for the case that the data are given exactly, and a proof that the discrepancy principle provides a regularizing stopping rule for the case of perturbed data.

Numerical examples (cf. [1]) show that the method is superior to CGLS and also to its preconditioned variant from [3] for space invariant deblurring problems, when using the same approximation C to set up the preconditioner in either case. Our results apply to different types of boundary conditions alike, including Dirichlet, reflective, and antireflective boundary conditions.

References

[1] M. Donatelli and M. Hanke, *Fast nonstationary preconditioned iterative methods for ill-posed problems, with application to image deblurring*, Inverse Problems **29** (2013) 095008.

[2] M. Hanke, *A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems **13** (1997), pp. 79–95.

[3] M. Hanke, J.G. Nagy, and R.J. Plemmons, *Preconditioned iterative regularization for ill-posed problems*, in L. Reichel, A. Ruttan, and R.S. Varga, editors, *Numerical Linear Algebra*, de Gruyter, Berlin, 1993, pp. 141–163.

[4] M. Hanke and J.G. Nagy, *Restoration of atmospherically blurred images by symmetric indefinite conjugate gradient techniques*, Inverse Problems **12** (1996), pp. 157–173.

[5] M.E. Kilmer, *Cauchy-like preconditioners for two-dimensional ill-posed problems*, SIAM J. Matrix Anal. Appl. **20** (1999), pp. 777–799.

Rotational Image Deblurring with Sparse Matrices

Per Christian Hansen, James G. Nagy and Konstantinos Tigkos

Abstract

Image deblurring is an important image restoration problem in its own right, as well as a component in navigation systems, robotics, etc. A common cause of blurring is motion of either the recording device or the object. Examples include movement of the heart during cardiac imaging processes, movement of a low flying aircraft in aerial imaging, and people or vehicle movement in surveillance imaging. Linear motion with constant speed and direction is well understood, but less work has been done in the development of mathematical models and efficient algorithms for restoration of images degraded by nonlinear and nonuniform motion blur where the blurring is spatially variant.

This work focuses on a particular type of motion blur where the object or scene is rotated during the recording of the image, leading to spatially variant blur. An important special case is when the rotation axis of the scene points towards the camera, and in this case, if the rotation also has constant velocity, then it is possible to transform the blurring to polar coordinates, which results in a shift-invariant operator. Although this is mathematically convenient, it is not possible to obtain a purely shift-invariant operator for more complicated, nonuniform motion blur. In addition, the transformation between rectangular and polar coordinates can introduce severe interpolation errors in the reconstruction.

We develop algorithms that avoid such transformations and thus are able to handle rotational blur along an arbitrary axis. Our algorithms are based on iterative algorithms [BN] that incorporate nonnegativity constraints, namely, the projected Landweber method and the modified residual norm steepest descent (MRNSD) method. Both algorithms exhibit semi-convergence, i.e., the early iterations produce iteration vectors that converge towards the desired noise-free solution, while later iterations produce noisy vectors that diverge from from this solution.

Algorithms for general motion blur typically use a matrix-free approach based on summing a sequence of rotated and/or translated images [TTB]. We develop an alternative approach in which we explicitly form a sparse matrix that represents the spatially variant rotational blur [HNT]. The matrix needs only be constructed once, its store requirements are moderate, and we can easily incorporate suitable boundary conditions into the matrix in order to reduce artifacts from the edges. The advantages are fast execution, at the expense of storage requirements for the sparse matrix, and flexible treatment of boundary conditions.

The performance of our algorithms is illustrated with numerical examples. We illustrate why motion deblurring tents to introduce replicas of sharp edges in the reconstruction. We we study the use of several stopping rules and conclude that the Monte-Carlo GCV method is the most robust one. And we demonstrate that the correct handling of boundary conditions is essential for achieving good reconstructions.

[BN] J. M. Bardsley and J. G. Nagy, *Covariance-preconditioned iterative methods for nonnegatively constrained astronomical imaging*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 1184–1197.

[HNT] P. C. Hansen, J. G. Nagy, and K. Tigkos, *Rotational image deblurring with sparse matrices*, submitted to BIT.

[TTB] Y. W. Tai, B. Tan, and M. Brown, *Richardson-Lucy deblurring for scenes under a projective motion path*, IEEE Trans. Pattern Analysis and Machine Intelligence, 33 (2011), pp. 1603–1618.

How and Why to Estimate Condition Numbers for Matrix Functions

Nicholas J. Higham, Lijing Lin and Samuel Relton

Abstract

The condition number of a matrix with respect to inversion was introduced by Turing in 1948. Since then various condition numbers have been defined in numerical linear algebra, with investigations focusing on characterizing when a problem is ill conditioned, efficiently estimating condition numbers, obtaining condition number bounds, and relating the level-2 condition number (the condition number of the condition number) to the original (level-1) condition number. In this talk we consider all these aspects for general matrix functions.

Consider a matrix function $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ and assume that it is Fréchet differentiable. The Fréchet derivative is a linear mapping $L_f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ such that for all $E \in \mathbb{C}^{n\times n}$

$$f(A + E) - f(A) - L_f(A, E) = o(\|E\|).$$

The Fréchet derivative describes the first-order sensitivity of f to perturbations. Since L_f is a linear operator,

$$\mathrm{vec}(L_f(A, E)) = K_f(A)\mathrm{vec}(E)$$

where $K_f(A) \in \mathbb{C}^{n^2\times n^2}$ is the Kronecker matrix and vec stacks the columns of its argument into one long vector. We will explain how the Kronecker matrix contains a wealth of information about sensitivity measured both normwise and componentwise.

Componentwise Sensitivity

First, we show that the $((s-1)n+r, (j-1)n+i)$ element of K_f is the absolute condition number of $f(A)_{rs}$ subject to perturbations in a_{ij}. Hence a norm of row $(s-1)n+r$ of $K_f(A)$ is an absolute condition number of $f(A)_{rs}$ subject to perturbations in A; equivalently, it is the norm of the gradient of the map from A to $f(A)_{rs}$. Similarly, a norm of column $(j-1)n+i$ of $K_f(A)$ is an absolute condition number of $f(A)$ subject to perturbations in a_{ij}. In practice we may wish to know the $k \ll n$ elements of $f(A)$ that are most sensitive to perturbations in (all the elements of) A or the k elements of A to which $f(A)$ is (overall) most sensitive to perturbations. We will show how the block 1-norm estimator of [5] can be used to estimate these elements and the corresponding sensitivities efficiently without explicitly forming the $n^2 \times n^2$ matrix $K_f(A)$.

Normwise Sensitivity

For a matrix function $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, the theory of Rice leads to absolute and relative normwise condition numbers characterized in terms of the norm of the Fréchet derivative of f. Estimating the condition number reduces to estimating the norm of the Kronecker matrix, We review how this can be done using the block 1-norm estimator. The estimator requires the evaluation of matrix–vector products involving the Kronecker matrix and its conjugate transpose. Obtaining the latter products is nontrivial in general. We show that these reduce to products with the Kronecker matrix for the function $\widetilde{f}(z) := \overline{f(\overline{z})}$ provided that $\widetilde{f}(A)^* = \widetilde{f}(A^*)$ for all $A \in \mathbb{C}^{n\times n}$ and that in most cases of interest $\widetilde{f} \equiv f$ and the latter condition holds.

We survey methods for computing or approximating the Fréchet derivative, including the complex step method for general f and methods specialized to particular f.

We then turn to the two questions:

Q1 What is the level-2 condition number of a matrix function and how does it relate to the (level-1) condition number?

Q2 What is the condition number of the Fréchet derivative and how can it be estimated?

Q1 was first raised in the context of problems including matrix inversion, and the eigenvalue problem by Demmel [1], who showed that for the problems in question the level-1 and level-2 condition numbers are equivalent. We briefly outline how to bound the level-2 condition number and describe several cases in which explicit relations between the level-1 and level-2 condition numbers can be obtained. More details are given in the separate abstract by Relton.

Q2 is important for understanding the behaviour of algorithms for computing the Fréchet derivative, as it tells us how large the relative error can be expected to be for a backward stable algorithm. We give an appropriate definition of condition number of the Fréchet derivative and show how to bound it and how to estimate the bound.

We conclude with a brief discussion of available software, including the Python package SciPy, which contains in version 0.13.0 several codes to compute Fréchet derivatives of matrix functions.

This talk is based on [2], [3], [4].

References

[1] James W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numer. Math.*, 51:251–289, 1987.

[2] Nicholas J. Higham and Lijing Lin. An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives. *SIAM J. Matrix Anal. Appl.*, 34(3):1341–1360, 2013.

[3] Nicholas J. Higham and Samuel D. Relton. The condition number of the Fréchet derivative of a matrix function. MIMS EPrint, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2013. In preparation.

[4] Nicholas J. Higham and Samuel D. Relton. Higher order Fréchet derivatives of matrix functions and the level-2 condition number. MIMS EPrint, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2013. In preparation.

[5] Nicholas J. Higham and Françoise Tisseur. A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra. *SIAM J. Matrix Anal. Appl.*, 21(4):1185–1201, 2000.

Noise Approximation in Discrete Ill-posed Problems

Iveta Hnětynková, Marie Michenková and Martin Plešinger

Abstract

In many fields of application, e.g. signal and image processing, geophysics, seismology, radiology, etc., there is a need to solve linear inverse problems $Ax \approx b$, where the matrix A represents a discretized smoothing operator, and b is an observation vector contaminated by unknown noise. By the nature of these problems they are typically ill-posed. A small perturbation in the data may result in significant errors in computed approximate solution; a naive solution is dominated by amplified noise. Thus it is necessary to use *regularization methods* for finding reliable numerical approximations to the solution. A knowledge of the noise level in the data is a great advantage as it allows to apply the Morozov's discrepancy principle [4].

In [2] it was shown, how noise contaminated in b propagates in the left bidiagonalization vectors of the Golub-Kahan iterative bidiagonalization. Analysis was based on quadrature approximations of the Rieman-Stieltjes distribution function associated with the given data. This allowed to *estimate the usually unknown noise level* at a negligible cost. The estimate required monitoring near stagnation of the absolute value of the first component of the left singular vector of the projected (bidiagonal) matrix corresponding to its smallest singular value. A possibility of *approximating the high frequency part of the unknown noise* was also mentioned. These results were presented at the previous Householder symposium in 2011.

In [2] the results were illustrated on examples from the Regularization Toolbox [1], and in order to maintain orthogonality among the bidiagonalization vectors close to the machine precision level, full (double) reorthogonalization was used. For real problems, the full reorthogonalization is too costly and the noise can be far from the artificially constructed "white" noise. We show that even if no reorthogonalization is used, the noise revealing effect is still present. However, we can observe appearance of multiple approximations for large singular values of A throughout the bidiagonalization. This causes reappearance of some smooth components in the left bidiagonalization vectors, with the propagation of noise delayed and sometimes irregular.

In [2], white noise was considered. If noise is high-frequency dominated, the revealing technique is by its nature expected to perform well. However, realistic noise can have more complicated properties. We investigate the behavior of the proposed noise revealing technique under various circumstances which could be met in real world applications.

Finally, we turn to approximation of the high frequency part of the unknown noise vector; see also [3]. We demonstrate that by subtracting such approximation from b the high frequency part of noise in the data is significantly reduced, but the corresponding part of the signal is also affected. Determining approximate noise in a reliable way requires further investigation; we hope for getting applicable results in the near future.

References

[1] Hansen, P. C.: Regularization Tools – version 3.2 for MATLAB 6.0, a package for analysis and solution of discrete ill-posed problems

[2] Hnětynková, I., Plešinger, M., Strakoš, Z.: The regularizing effect of the Golub-Kahan iterative bidiagonalization and revealing the noise level in the data, BIT **49**, pp. 669–696 (2009)

[3] Michenková, M.: Regularization techniques based on the least squares method, diploma thesis, Charles University in Prague (2013)

[4] Morozov, V. A.: On the solution of functional equations by the method of regularization (in Russian), Soviet Math. Dokl. **7**, pp. 414–417 (1966)

Field of Values type Eigenvalue Inclusion Regions for Large Matrices

Michiel Hochstenbach and Ian N. Zwaan

Abstract

We will discuss several recent contributions in the development of tight and very fast spectral inclusion regions for large sparse matrices, based on the field of values

$$W(A) = \{\, \mathbf{x}^* A \mathbf{x} \;:\; \|\mathbf{x}\|_2 = 1 \,\}$$

and generalizations.

- For some matrices, such as matrices of the tolosa family, the field of values turns out to be much larger than the spectrum. We will propose various **(Krylov) scaling techniques** for this situation, showing that scaling of a matrix may be a very helpful technique for generating tight spectral inclusion regions based on a field of values.

  In fact, we believe that the combination of matrix scaling and a field of values based on an Arnoldi decomposition gives an eigenvalue inclusion region that is very hard to beat both in quality and efficiency.

- We will discuss **adapted fields of values** for the situation that the matrix has some **outlier eigenvalues**. In this case, the field of values would be a spectral inclusion region that would be much larger than necessary for the bulk of the eigenvalues.

- Finally, we will also discuss eigenvalue inclusion regions based on the field of values for the **generalized eigenvalue problem** (matrix pencils) and the **quadratic eigenvalue problem**.

Indeed, we would like to stress the almost astonishing result that quality eigenvalue inclusion regions for large sparse matrices may be obtained with just a dozen matrix-vector products. This is surprising since finding just one eigenvalue accurately may cost hundreds or even thousands of matrix-vector products.

Part of this presentation is described in [2, 5]. Some earlier relevant work was done in [1, 3, 4].

References

[1] T.-Y. Chen and J. W. Demmel, *Balancing sparse matrices for computing eigenvalues*, Linear Algebra and Its Applications, 309 (2000), pp. 261–287.

[2] M. E. Hochstenbach, *Fields of values and inclusion regions for matrix pencils*, Electron. Trans. Numer. Anal., 38 (2011), pp. 98–112.

[3] M. E. Hochstenbach, D. A. Singer, and P. F. Zachlin, *Eigenvalue inclusion regions from inverses of shifted matrices*, Linear Algebra Appl., 429 (2008), pp. 2481–2496.

[4] ——, *Numerical approximation of the field of values of the inverse of a large matrix*, Textos de Matematica, 44 (2013), pp. 59–71.

[5] M. E. Hochstenbach and I. N. Zwaan, *Matrix balancing for field of values type inclusion regions*, submitted.

A Schur Logarithmic Algorithm for Fractional Powers of Matrices

Bruno Iannazzo and Carlo Manasse

Abstract

A fractional power of a square matrix A is any solution of the matrix equation $X^p = A^q$, where p and q are positive integers.

Fractional powers can be classified either primary or nonprimary, whether or not they can be written as a polynomial of A. Perhaps surprising, in the generic case, every fractional power is primary. More precisely, for a nonsingular matrix A, nonprimary powers arise if only if A has at least two Jordan blocks for the same eigenvalue in its Jordan canonical form; these nonprimary powers are non-isolated points in the space of matrices and are ill-conditioned, for this reason, numerical algorithms focus just on primary powers.

For matrices having no nonpositive real eigenvalues, the fractional power $\exp(\frac{q}{p}\log(A))$ is said to be principal and it is the one usually required in applications.

In fact, the problem of computing fractional powers of matrices arises in several applications, of which we cite just a couple. In finance, credit rating agencies produce transition matrices whose entries are the probabilities to skip from a rate to another in a fixed period. If one is interested in these probabilities, but over a shorter period, then a fractional power may be helpful [6]. As usual in applied mathematics, the same model can be used for a very different problem: in the study of chronic degenerative diseases, the entries of the transition matrix represent the probability of transition from a state of the disease to another one; fractional powers are useful to describe the transition over a short period, given the transition observed in a long period [1].

Notice that a stochastic matrix whose principal pth root is not stochastic may still have a primary stochastic pth root [3] and thus one might be interested also in computing non-principal fractional powers.

We present an algorithm for computing all primary fractional powers of a nonsingular matrix [5]. The algorithm belongs to the category of Schur recurrence algorithms, which compute a root of a matrix using a suitable recurrence on the Schur form of A. The nice feature of the proposed algorithm is that, in terms of p and the size n of the matrix A, it requires $O(n^3 \log p)$ arithmetic operations, which reduces the cost with respect to the previous Schur recurrence algorithms [2, 7]. The algorithm works completely in real arithmetic for real data and shows excellent numerical behavior, in the case $q = 1$. Moreover, for moderate values of p its computational cost is competitive with existing algorithms for the principal fractional powers [4].

We consider also the related problem of computing Fréchet derivatives of fractional powers functions, which are useful to get the condition number of the fractional power in which we are interested. In fact, a primary fractional powers of a nonsingular matrix A can be uniquely extended in a neighborhood of A to a differentiable function $f(Z)$ verifying the equation $(f(Z))^p = Z^q$, for any Z in the mentioned neighborhood.

Finally, we consider some implementation issues related to the possibility to use the modern parallel architectures to increase the performance of the proposed algorithms.

References

[1] T. Charitos, P. R. de Waal and L. C. van der Gaag, *Computing short-interval transition matrices of a discrete-time Markov chain from partially observed data*, Stat. Med. 27-6 (2008), pp. 905–921.

[2] F. Greco and B. Iannazzo, *A binary powering Schur algorithm for computing primary matrix roots*, Numer. Algorithms, 55 (2010), pp. 59–78.

[3] N. J. Higham and L. Lin, *On pth roots of stochastic matrices*, Linear Algebra Appl., 435 (2011), pp. 448–463.

[4] N. J. Higham and L. Lin, *A Schur-Padé algorithm for fractional powers of a matrix*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1056–1078.

[5] B. Iannazzo and C. Manasse, *A Schur logarithmic algorithm for fractional powers of matrices*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 794–813.

[6] R. B. Israel, J. S. Rosenthal and J. Z. Wei, *Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings*, Math. Finance 11-2 (2001), pp. 245–265.

[7] M. I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.

Randomized Algorithms for Numerical Linear Algebra

Ilse Ipsen

Abstract

Randomized algorithms for matrix computations are starting to gain ground for computations with massive datasets in application areas like nuclear engineering, population genomics, and astronomy. These computations include matrix multiplication; least squares and regression problems; low rank approximation and dimensionality reduction (PCA, CUR, subset selection); and preconditioning methods. We will analyze the error due to randomization and the sensitivity to perturbations; and present a Matlab toolbox designed for evaluating randomized sampling methods and probabilistic bounds.

The Infinite Arnoldi Method for the Waveguide Eigenvalue Problem

Elias Jarlebring and Olof Runborg

Abstract

Consider the propagation of (time-harmonic) waves in a periodic medium with certain properties. The wave modes can in this situation be described with the quadratic PDE eigenvalue problem on an infinite strip,

$$\begin{aligned} \Delta u + 2\lambda u_z + (\lambda^2 + \kappa^2)u = 0, &\qquad (x,z) \in \mathbb{R} \times (0,1), \\ u(x,0) = u(x,1) &\qquad x \in \mathbb{R}, \\ \lim_{|x|\to\infty} u(x,z) = 0, &\qquad z \in [0,1], \end{aligned}$$

where $\kappa = \kappa(x,z)$ is a given function and $\lambda \in \mathbb{C}$ is the eigenvalue. We present a new approach for this PDE eigenvalue problem. The application of certain discretization techniques leads to a nonlinear eigenvalue problem with a particular structure. A new algorithm is derived which exploits the structure of the nonlinearity as well as the matrices arising in the discretization. This PDE eigenvalue problem is common in the study of waveguides. See [1] and references therein for literature on waveguides.

More precisely, we first transform the problem to a finite domain $[x_-, x_+] \times [0,1]$ by using artificial boundary conditions, in particular so-called Dirichlet-to-Neumann maps. These artificial boundary conditions depend on the eigenvalue and a discretization consequently leads to a nonlinear eigenvalue problem: Find $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n \backslash \{0\}$ such that

$$M(\lambda)v = 0, \tag{1}$$

where M is a holomorphic function. See, e.g., [2, 3, 4] for literature on nonlinear eigenvalue problems of this type. In our situation, we have explicitly

$$M(\lambda) := \begin{pmatrix} A_0 + A_1\lambda + A_2\lambda^2 & C_1 \\ C_2^T & \alpha I + R^H \Lambda(\lambda) R \end{pmatrix} \tag{2}$$

and A_0, A_1, A_2 are large and sparse matrices corresponding to the discretization of the interior domain. The matrix $\Lambda(\lambda) \in \mathbb{C}^{n_x \times n_x}$ is a diagonal matrix with entries $\pm i\sqrt{-4\pi^2k^2 + 4i\pi k\lambda + \lambda^2 + \kappa_\pm^2}$ for integer k and $\kappa_\pm \in \mathbb{R}_+$. The matrix R is a Vandermonde matrix, which never will be formed explicitly since the action can be computed with the fast Fourier transform (FFT).

We adapt the algorithm called the infinite Arnoldi method [5] for this problem. The infinite Arnoldi method is an algorithm based on an infinite-dimensional operator, whose (reciprocal) eigenvalues are equivalent to the solutions of (1) when M is holomorphic. The method is equivalent to Arnoldi's method applied to the infinite-dimensional operator, where the structure of the operator is exploited such that it can be carried out with finite-dimensional operations. We specialize the infinite Arnoldi method as follows.

- The matrix M has branch-point singularities. They are unfortunately close to several of the eigenvalues of interest and limit the convergence and reliability of the direct application of the infinite Arnoldi method. We characterize these points and resolve the issue by carrying out a Cayley transformation $\gamma = (\lambda - \lambda_0)/(\lambda + \overline{\lambda_0})$ and restating a nonlinear eigenvalue problem in γ. The reformulated nonlinear eigenvalue problem has branch points at more favorable locations.

- The infinite Arnoldi method requires that certain quantities associated with M can be accurately and efficiently computed. We show how these quantities can be computed for the Cayley transformed problem. It turns out that we can explicitly compute the quantities by using functions of matrices (the matrix square root). We also use that the action of R and R^H can be formed with FFT and inverse FFT.

- After explicitly taking the structure of M into account, the computationally dominating part of the infinite Arnoldi method applied to this problem is the orthogonalization. Due to a growth in the basis matrix, the computational cost associated with the orthogonalization when carrying out k steps of the infinite Arnoldi method is $O(k^3)$. We show that the basis matrix has a particular structure which can be exploited by representing it in a factorized form using tensors. The orthogonalization can be carried out on the factorized representation directly and reduces the complexity of the orthogonalization cost to $O(k^2)$. The tensor-representation technique is general and should be useful also for other large-scale sparse nonlinear eigenvalue problems.

We illustrate the properties of the algorithm with simulations. In particular, the efficiency and reliability of the algorithm is illustrated with a model of a waveguide. We show that the algorithm accurately and efficiently solves the waveguide eigenvalue problem for a periodic medium with complicated geometry, which cannot be easily treated with alternative methods.

References

[1] J. Tausch, J. Butler, Floquet multipliers of periodic waveguides via Dirichlet-to-Neumann maps, J. Comput. Phys. 159 (1) (2000) 90–102.

[2] V. Mehrmann, H. Voss, Nonlinear eigenvalue problems: A challenge for modern eigenvalue methods, GAMM Mitteilungen 27 (2004) 121–152.

[3] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, F. Tisseur, NLEVP: A collection of nonlinear eigenvalue problems, ACM Transactions on Mathematical Software 39 (2) (2013) 1–7.

[4] C. Effenberger, Robust solution methods for nonlinear eigenvalue problems, Ph.D. thesis, EPF Lausanne (2013).

[5] E. Jarlebring, W. Michiels, K. Meerbergen, A linear eigenvalue algorithm for the nonlinear eigenvalue problem, Numer. Math. 122 (1) (2012) 169–195.

The Geometric Matrix Mean: an Adaptation for Structured Matrices

Dario A. Bini, Bruno Iannazzo, Ben Jeuris and Raf Vandebril

Abstract

Positive definite matrices can be encountered in a widespread collection of applications, such as signal processing, bio-informatics, and radar technology. As a consequence, these matrices and their geometry has been well researched, resulting in a natural and optimal geometry.

In many applications, not one, but multiple positive definite matrices are provided through measurements. It is desired to find an average representation of the matrices, which best describes the corresponding measurement. To this end, the geometric matrix mean is often used, because it has interesting properties with regard to positive definite matrices, such as invariance under inversion, invariance under congruence, etc. Of the various instances for the geometric mean, the Karcher mean appears the most natural through its definition as the barycenter of the matrices under the positive definite geometry. Computationally, the barycenter is located using matrix manifold optimization techniques.

Often positive definiteness is not the only structural property present in the measured matrices. Additional structures can be, e. g., Toeplitz or Hankel structure, low displament rank, etc. These structures are usually linked to some physical interpretation in the application, so it is desirable for the result to preserve the structure and hence the interpretation. Unfortunately, the Karcher mean is typically not structure preserving.

We present an adaptation of the Karcher mean which preserves desired structures by computing an optimizer of the barycenter problem over a restricted search space. The interesting properties of the geometric mean are generalized to take the additional structure into account.

To compute the new barycenter, two preconditioned gradient descent methods are investigated for linear submanifolds of the positive definite matrix manifold. The preconditioners are derived using differential geometry, where the natural positive definite geometry results in a more involved, but also more efficient preconditioner.

In a second approach, we focus on the set of positive definite Toeplitz matrices. An application-inspired transformation maps such a matrix to the product space of the positive scalars and a multiple of the complex unit circle. Separately, these sets are naturally endowed with the geometry of positive scalars (the one-dimensional equivalent of the positive definite geometry) and the hyperbolic geometry, respectively. Combined, this allows the definition of a new barycenter.

Finally, we test a generalization of the previous barycenter to the set of Block-Toeplitz Toeplitz-block (positive definite) matrices. The new blockwise transformation generalises the positive scalars to positive definite matrices and provides a natural extension of the hyperbolic unit circle to the matrix setting. The final algorithms are again designed to maintain the original structure of the matrices.

The performance of all mentioned algorithms is tested in numerical experiments, both in a theoretical environment and in real-world applications. The unstructured Karcher mean is applied in an experiment in bio-informatics where correlation between human genes and physical traits is combined over several independent measurements. For the structured versions, an application is found in radar detection, where positive definite Toeplitz matrices need to be averaged.

Stratification of some Structured Matrix pencil Problems: how Canonical Forms Change under Perturbations

B. Kågström

Abstract

In this presentation, we will review and highlight some of our recent results concerning the stratification of orbits (and bundles) of structured matrix pencil problems, focusing on (i) companion form linearizations of full normal-rank polynomial matrices and (ii) general skew-symmetric matrix pencils.

The stratification theory for matrix pencils provides the complete picture of nearby canonical structures and how structure transitions take place under perturbations. More formally, it reveals the qualitative information that the closure hierarchy of orbits and bundles provides (e.g., see [4, 5, 6, 8] and references therein). For example, the equivalence orbit of a matrix pencil $A - \lambda B$ consists of all pencils with the same eigenvalues and the same canonical form under strict equivalence transformations $U^{-1}(A - \lambda B)V$, where U and V are non-singular. A bundle is the union of all orbits with the same canonical form but with unspecified eigenvalues. The closure hierarchy is determined by the closure and cover relations among orbits (or bundles), where a cover relation guarantees that two orbits (or bundles) are nearest neighbours in the hierarchy. In a stratification, an orbit can never be covered by a less or equally generic orbit. This means that orbits (canonical forms) within the closure hierarchy can be ordered by their dimension (or codimension) [1].

Polynomial matrices play an important role in the study of dynamical systems described by sets of differential-algebraic equations (DAEs) with constant coefficient matrices. Here, we consider *full rank polynomial matrices*

$$P(s) := P_d s^d + \ldots + P_1 s + P_0, \ s \in \mathbb{C},$$

where the leading coefficient matrix P_d is nonzero so that the highest degree is indeed d (we say it has *exact degree d*). The eigenstructure elements of a polynomial matrix are defined via the Smith normal form under unimodular transformations. The classical approach to analyze and determine the structural elements of $P(s)$ is to study linearizations. In [7], we show that perturbations of polynomial matrices of full normal-rank can be analyzed via the study of perturbations of *companion form linearizations* of such polynomial matrices. It is proved that a full normal-rank polynomial matrix has the same structural elements as its right (or left) linearization. Furthermore, the linearized pencil has a *special block matrix structure* that can be taken into account when studying its stratification. This yields constraints on the set of achievable eigenstructures, which allowed us to derive necessary and sufficient conditions for cover relations between two orbits or bundles of the linearization of full normal-rank polynomial matrices. Besides, reviewing these results, we will illustrate the stratification rules applied to a half-car passive suspension system with four degrees of freedom.

In the second part, we address *skew-symmetric matrix pencils* $A - \lambda B$, where $A^T = -A$ and $B^T = -B$, by studying how small perturbations may change their canonical forms under structure-preserving *congruence transformations* $V^T(A - \lambda B)V$ with V non-singular. In general, reduction of a complex skew-symmetric matrix pencil $A - \lambda B$ to any canonical form under congruence is an unstable operation: both the canonical form and the reduction transformation depend discontinuously on the entries of $A - \lambda B$. Thus it is important to know the canonical forms of all such pencils that are arbitrarily close to $A - \lambda B$. We solve this problem by deriving the *closure hierarchy graphs* (i.e., stratifications) of orbits and bundles of skew-symmetric matrix pencils [2, 3].

The set of all matrix pencils strictly equivalent to a skew-symmetric matrix pencil $A - \lambda B$ as well as the set of matrix pencils congruent to $A - \lambda B$ are the orbits of $A - \lambda B$ under the actions of strict equivalence ($O^e_{A-\lambda B}$) and congruence ($O^c_{A-\lambda B}$), respectively. Note that $O^e_{A-\lambda B}$ contains non-skew-symmetric pencils too. We prove that for two skew-symmetric matrix pencils $A - \lambda B$ and $C - \lambda D$ the equivalence orbit inclusion $\overline{O^e_{C-\lambda D}} \supset O^e_{A-\lambda B}$ is equivalent to the congruence orbit inclusion $\overline{O^c_{C-\lambda D}} \supset O^c_{A-\lambda B}$. In other words, we generalize the fact that two skew-symmetric matrix pencils are strictly equivalent if and only if they are congruent, proving that a skew-symmetric matrix pencil $A - \lambda B$ can be approximated by pencils strictly equivalent to a skew-symmetric matrix pencil $C - \lambda D$ if and only if $A - \lambda B$ can be approximated by pencils congruent to $C - \lambda D$.

The results presented are collaborative efforts together with several people including Andrii Dmytryshyn, Stefan Johansson, Vladimir Sergeichuk, and Paul Van Dooren.

References

[1] A. Dmytryshyn, S. Johansson, and B. Kågström. *Codimension computations of congruence orbits of matrices, symmetric and skew-symmetric matrix pencils using Matlab*. Report UMINF 13.18, Department of Computing Science, Umeå University, 2013.

[2] A. Dmytryshyn and B. Kågström, On closure hierarchies of skew-symmetric matrix pencils. *Manuscript*, 2013.

[3] A. Dmytryshyn, B. Kågström, V.V. Sergeichuk, Skew-symmetric matrix pencils: codimension counts and the solution of a pair of matrix equations. *Linear Algebra Appl.*, 438:3375–3396, 2013.

[4] A. Edelman, E. Elmroth, and B. Kågström. A Geometric Approach to Perturbation Theory of Matrices and Matrix Pencils. Part I: Versal Deformations. *SIAM J. Matrix Anal. Appl.*, 18(3):653–692,1997. Awarded **the SIAM Linear Algebra Prize 2000**.

[5] A. Edelman, E. Elmroth, and B. Kågström. A Geometric Approach To Perturbation Theory of Matrices and Matrix Pencils. Part II: A Stratification-Enhanced Staircase Algorithm. *SIAM J. Matrix Anal. Appl.*, 20:667–699, 1999.

[6] E. Elmroth, S. Johansson, and B. Kågström. Stratification of controllability and observability pairs — Theory and use in applications. *SIAM J. Matrix Anal. Appl.*, 31(2):203–226, 2009.

[7] S. Johansson, B. Kågström, and P. Van Dooren. Stratification of full rank polynomial matrices. *Linear Algebra Appl.*, 439:1062–1090, 2013.

[8] B. Kågström, S. Johansson, and P. Johansson. StratiGraph Tool: Matrix Stratification in Control Applications. In L. Biegler, S. L. Campbell, and V. Mehrmann, editors, *Control and Optimization with Differential-Algebraic Constraints*, chapter 5. SIAM Publications, 2012. ISBN 978-1-611972-24-5.

Fast Generation of Random Orthogonal Matrices

Nicholas J. Higham, Amal Khabou and Françoise Tisseur

Abstract

Random orthogonal matrices have a wide variety of applications. They are used in the generation of various kinds of random matrices and random matrix polynomials [1], [2], [3], [7]. They are used to randomize integration methods for n-dimensional integrals over spherically symmetric integration regions [5]. The random orthogonal matrix (ROM) simulation [8] method uses random orthogonal matrices to generate multivariate random samples with the same mean and covariance as an observed sample.

The natural distribution over the space of orthogonal matrices is the Haar distribution. One way to generate a random orthogonal matrix from the Haar distribution is to generate a random matrix A with elements from the standard normal distribution and compute its QR factorization $A = QR$, where R is chosen to have nonnegative diagonal elements; the orthogonal factor Q is then the required matrix [6].

Stewart [10] develops a more efficient algorithm that directly generates an $n \times n$ orthogonal matrix from the Haar distribution as a product of Householder transformations built from Householder vectors of dimensions $1, 2, \ldots, n-1$ chosen from the standard normal distribution. This algorithm is implemented in the LAPACK test software [3] and in MATLAB as `gallery('qmult')`. The LAPACK test software and the MATLAB function call `gallery('randsvd')` both generate $m \times n$ matrices of the form $A = U\Sigma V^*$ where U $(m \times m)$ and V $(n \times n)$ are from the Haar distribution and Σ is a diagonal matrix containing a specified set of singular values. Stewart's algorithm requires $m^3 + n^3$ flops to form A.

The goal of this work is to design an algorithm that significantly reduces the computational cost of Stewart's algorithm by giving up the property that Q is exactly Haar distributed.

We generate orthogonal random Q of the form $Q = HDH^*$, where $H = H_1 H_2 \ldots H_k$ is a product of Householder transformations based on full vectors with elements from the standard normal distribution and D is block diagonal with 2×2 and 1×1 real diagonal blocks. The diagonal blocks are real and are chosen to have eigenvalues with phases uniformly distributed on the unit circle, since this is a property of the Haar distribution [4], [9].

We will argue that for the purpose of test matrix generation we can take k much less than the matrix dimension and still obtain an acceptable matrix. Because each H_j is a full matrix the product defining Q can be computed more efficiently (per flop) than with Stewart's algorithm, in which most operations are effectively on matrices of dimension less than n.

We will give performance results on a variety of architectures to show the benefits of the new algorithm.

References

[1] Timo Betcke, Nicholas J. Higham, Volker Mehrmann, Christian Schröder, and Françoise Tisseur, *NLEVP: A collection of nonlinear eigenvalue problems*, ACM Trans. Math. Software, 39 (2013), pp. 7:1–7:28.

[2] Philip I. Davies and Nicholas J. Higham, *Numerically stable generation of correlation matrices and their factors*, BIT, 40 (2000), pp. 640–651.

[3] James W. Demmel and A. McKenney, *A test matrix generation suite*, Preprint MCS-P69-0389, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA, Mar. 1989. LAPACK Working Note 9.

[4] Alan Edelman and N. Raj Rao, *Random matrix theory*, Acta Numerica, 14 (2005), pp. 233–297.

[5] Alan Genz, *Methods for generating random orthogonal matrices*, proceedings of Monte Carlo and Quasi-Monte Carlo Methods 1998, H. Niederreiter and J. Spanier, eds., Springer-Verlag, (Berlin 2000), pp. 199–213.

[6] Richard M. Heiberger, *Algorithm AS 127: Generation of random orthogonal matrices*, J. Roy. Statist. Soc. Ser. C (Applied Statistics), 27 (1978), pp. 199–206.

[7] Nicholas J. Higham, *J-orthogonal matrices: Properties and generation*, SIAM Rev., 45 (2003), pp. 504–519.

[8] Walter Ledermann, Carol Alexander, and Daniel Ledermann, *Random orthogonal matrix simulation*, Linear Algebra Appl., 434 (2011), pp. 1444–1467.

[9] Francesco Mezzardi, *How to generate random matrices from the classical compact groups*, Notices Amer. Math. Soc., 54 (2007), pp. 592–604.

[10] G. W. Stewart, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.

Model Correction using a Nuclear Norm Constraint

Ning Hao, Lior Horesh and Misha E. Kilmer

Abstract

Let $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ be a comprehensive observation operator which transforms input $x \in \mathbb{R}^n$ into the observable space. Let $d \in \mathbb{R}^m$ be an observation obtained through the following relation:

$$d = \mathcal{F}(x) + \epsilon \tag{1}$$

where ϵ stands for measurement noise.

In many real world applications, it is often the case that our ability to prescribe $\mathcal{F}$ fully is limited. For example, in partial blind deconvolution, we may only know an approximation to the blurring operator. In other applications, we may only have access to (or can afford to run) simulations with a low resolution or low fidelity simulator. Thus, $\mathcal{F}$ is only partially specified, and hence the model which assumes this inaccurate operator is *misspecified*. Our goal is to formulate and solve a **model correction optimization problem** that will allow us to recover the true observation operator, assuming we are given a set of simulation model input parameters and a corresponding set of high fidelity output data. Such data can be obtained in various ways, such as through experimentation with known input models, analytic derivation, or alternatively through the use of a computationally intensive high-fidelity simulation.

Specifically, we formulate the model correction problem as a constrained stochastic optimization problem [5], where the objective function to be optimized includes a measure of discrepancy between the expected output of the current (low-fidelity) model along with the unknown correction against the data.

In this talk, we will assume an additive model of the form

$$\mathcal{F}(x) = A(x) + B(x), \tag{2}$$

where A is known, B needs to be estimated, and A and B are maps that take x to the data space. The model specification, or correction, problem is to recover B. We will need to set additional constraints on B to try to obtain a well-posed optimization problem.

We propose a Sample Average Approximation approach (SAA) [1]. Our general framework is to solve

$$\hat{B} = \arg\min_B \frac{1}{n_x n_\epsilon} \sum_{i=1,j=1}^{n_x,n_\epsilon} \mathcal{D}\left(A(x_i) + B(x_i), d_{i,j}\right),$$

subject to a structural constraint on B, and $\mathcal{D}$ defines our metric.

In this talk, we will limit our discussion to the case where B is linear (though A may not be linear) and the objective function is quadratic in B. We focus on the case where the desired structural constraint we wish to impose on B is that it be low rank. One motivation for such a constraint is in the context of recovering part of the signal subspace in a discrete ill-posed problem when only a rank-deficient operator is known (i.e. A is linear and rank deficient such that its range contains only the smoothest modes of the forward operator). We will attempt to enforce the low rank condition by using a nuclear norm constraint. We show that to solve this formulation of the problem, we can leverage Jaggi and Sulovskỳ's [2] scaling strategy applied to Hazan's algorithm [3] to incorporate minimization of the nuclear norm. This algorithm exhibits the nice property of requiring only an

(inexact) approximation to the largest eigenpair of a (symmetric) matrix that is derived from first order derivatives corresponding to the original $\mathcal{D}$. The algorithm also features a mechanism for keeping control over the rank of the recovered B. We give examples from partially-blind image deconvolution that demonstrate the potential of our approach.

We extend our model correction framework and algorithm to the case that A and B represent abstractions of operators. Specifically, we consider the case that B is a tensor representation of a desired correction, and focus on adapting our optimization problem to handle a tensor nuclear norm [4] constraint.

References

[1] A. Shapiro and D. Dentcheva and A. Ruszczyński, *Lecture Notes on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2009.

[2] M. Jaggi and M. Sulovsky, "A Simple Algorithm for Nuclear Norm Regularized Problems," *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[3] E. Hazan, "Sparse approximate solutions to semidefinite programs," *LATIN*, pp 306 - 316, 2008.

[4] N. Hao, "Moving from Matrix to Tensor-based Analysis and Algorithms for Applications in Imaging Science and Beyond," Ph.D. Thesis, Tufts University, 2013.

[5] N. Hao, L. Horesh, M. E. Kilmer, "Nuclear Norm Optimization and its Application to Observation Model Specification," in *Compressed Sensing and Sparse Filtering*, Springer Series on Signals and Communication Techonology, A. Y. Carmi, L. S. Mihaylova, S. J. Godsill, eds., 2013.

Numerical Linear Algebra and Matrix Theory in Action

Andrew Knyazev

Abstract

Numerical linear algebra and matrix theory often play central roles in computations involving optimization, differential equations, signal and image processing, and control. In this talk, we present such examples of recent projects at http://www.merl.com/ covering these application areas.

Computing Linear Combinations of φ Functions

Antti Koskela and Alexander Ostermann

Abstract

We consider a new Krylov subspace algorithm for computing expressions of the form

$$\sum_{\ell=0}^{p} h^\ell \varphi_\ell(hA) w_\ell, \tag{1}$$

where $A \in \mathbb{C}^{n\times n}$, $w_\ell \in \mathbb{C}^n$. The functions φ_ℓ, defined by

$$\varphi_\ell(z) = \sum_{j=0}^{\infty} \frac{z^j}{(j+\ell)!} = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{\mathrm{e}^\lambda}{\lambda^\ell} \frac{1}{\lambda - z}\, \mathrm{d}\lambda,$$

where Γ encircles z and 0, are related to the exponential function. Computational problems of this form appear when applying exponential integrators to large dimensional ODEs in semilinear form $u'(t) = Au(t) + g(u(t))$ (see [2, 4]).

The sum (1) can be computed using a single product of the exponential of an augmented matrix and a vector (see [1]). As an alternative, we present a Krylov iteration, which is based on the observation that (1) can be expressed as a Taylor series

$$u(h) = \sum_{\nu=0}^{\infty} \frac{h^\nu}{\nu!} m_\nu \tag{2}$$

with moments $m_\nu = \sum_{\ell=0}^{\min(\nu,p)} A^{\nu-\ell} w_\ell$ that satisfy the recursions

$$\begin{aligned} m_0 &= w_0, \\ m_\nu &= A m_{\nu-1} + w_\nu, && \text{for } 1 \le \nu \le p, \\ m_\nu &= A m_{\nu-1}, && \text{for } \nu > p. \end{aligned}$$

From (2) one can see that

$$u(h) \in \operatorname{span}\{m_0, m_1, \ldots\}.$$

We perform an Arnoldi-like iteration to obtain an orthonormal basis $Q_k \in \mathbb{C}^{n\times k}$ for the subspace

$$\operatorname{span}\{m_0, m_1, \ldots, m_{k-1}\}$$

and approximate $u(h)$ by using the projection of A onto to this subspace

$$F_k = Q_k^* A Q_k.$$

Setting

$$W = [w_p, \ldots, w_1], \quad V = [v_p, \ldots, v_1] = Q_k^* W, \quad \text{and} \quad v_0 = Q_k^* w_0,$$

the approximation $u_k(h) \approx u(h)$ is given by

$$u_k(h) = Q_k v_k(h),$$

where

$$v_k(h) = \sum_{\ell=0}^{p} h^\ell \varphi_\ell(hF_k) v_\ell.$$

By using the QR decomposition

$$[m_0, m_1, \ldots, m_{k-1}] = Q_k R_k,$$

we construct the matrix $F_k \in \mathbb{C}^{k\times k}$ such that the following Arnoldi-like relation holds

$$AQ_k = Q_k F_k - (I - Q_k Q_k^*)\,\widehat{W}\,R_k^{-1} + f_{k+1,k} q_{k+1} e_k^{\mathsf{T}},$$

where $\widehat{W} = \begin{bmatrix} w_1 & \ldots & w_j & 0 \end{bmatrix} \in \mathbb{C}^{n\times k}$, with $j = \min(k, p)$. This results in an algorithm for updating the basis Q_k.

Using Cauchy's integral formula we derive an a priori error bound which gives insight into the convergence behaviour of the algorithm. Also an efficient a posteriori error estimate is derived.

One interesting application for φ functions comes from classical electromagnetism. We consider the equations of motion for a single particle given by

$$\begin{aligned} \frac{d}{dt} q &= p/m, \\ \frac{d}{dt} p &= E(q) + \Omega(q) p, \end{aligned}$$

where $E(q)$ represents an external magnetic field, and $\Omega(q)$ is a skew-symmetric 3×3 matrix representing the Lorentz forces from an external magnetic field. For example, in case of a constant magnetic field $\Omega(q) \equiv \Omega$, an exponential version of the velocity Verlet algorithm is given by

$$\begin{aligned} q_1 &= q_0 + \frac{h}{m}\varphi_1(h\Omega) p_0 + \frac{h^2}{m}\varphi_2(h\Omega) E(q_0), \\ p_1 &= \exp(h\Omega) p_0 + h\varphi_1(h\Omega)(E(q_1) + E(q_0))/2. \end{aligned}$$

Here, $\exp(h\Omega)$ and the φ functions are easily computed with the help of the Rodrigues' formula. By using splitting strategies we construct integrators that preserve well the structure of the equation (energy, symmetry, volume).

References

[1] A.H. Al-Mohy and N.J. Higham, *Computing the action of the matrix exponential, with an application to exponential integrators*, SIAM J. Sci. Comput., 33 (2010), pp. 488–511.

[2] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numerica 19 (2010), pp. 209–286.

[3] C. Knapp, A. Koskela, A. Ostermann, and A. Kendl, *Exponential splitting methods for plasma physics.*, In preparation.

[4] A.Koskela and A.Ostermann, *Exponential Taylor methods: analysis and implementation*, Comput. Math. Appl., 65 (2013), pp. 487–599.

[5] A.Koskela and A.Ostermann, *A moment-matching Arnoldi iteration for linear combinations of φ functions*, Submitted for publication.

On the Convergence of the Residual Inverse Iteration for Nonlinear Eigenvalue Problems

Cedric Effenberger and Daniel Kressner

Abstract

We consider nonlinear eigenvalue problems of the form

$$T(\lambda)x = 0, \qquad x \neq 0, \tag{1}$$

where $T : D \to \mathbb{C}^{n\times n}$ is a continuously differentiable matrix-valued function on some open interval $D \subset \mathbb{R}$.

In the following, $T(\lambda)$ is supposed to be Hermitian for every $\lambda \in D$. Moreover, we assume that the scalar nonlinear equation

$$x^*T(\lambda)x = 0 \tag{2}$$

admits a unique solution $\lambda \in D$ for every vector x in an open set $D_\rho \subset \mathbb{C}^n$. The resulting function $\rho : D_\rho \to D$, which maps x to the solution λ of (2), is called *Rayleigh functional*, for which we additionally assume that

$$x^*T'(\rho(x))x > 0 \qquad \forall x \in D_\rho.$$

The existence of such a Rayleigh functional entails a number of important properties for the eigenvalue problem (1), see [4, Sec. 115.2] for an overview. In particular, the eigenvalues in D are characterized by a min-max principle and thus admit a natural ordering. Specifically, if

$$\lambda_1 := \inf_{x\in D_\rho} \rho(x) \in D$$

then λ_1 is the first eigenvalue of T.

In this paper, we study the convergence of Neumaier's residual inverse iteration [3] for computing the eigenvalue λ_1 of T and an associated eigenvector x_1. In the Hermitian case, this iteration takes the form

$$v_{k+1} = \gamma_k(v_k + P^{-1}T(\rho(v_k))v_k), \quad k = 0, 1, \ldots, \tag{3}$$

with normalization coefficients $\gamma_k \in \mathbb{C}$, an initial guess $v_0 \in \mathbb{C}^n$, and a Hermitian preconditioner $P \in \mathbb{C}^{n\times n}$. Usually, $P = -T(\sigma)$ for some shift σ not too far away from λ_1 but the general formulation (3) allows for more flexibility, such as the use of multigrid preconditioners.

In [3, Sec. 3], it was shown that (3) with $P = T(\sigma)$ converges linearly to an eigenvector belonging to a simple eigenvalue, provided that σ is sufficiently close to that eigenvalue. Jarlebring and Michiels [2] derived explicit expressions for the convergence rate by viewing (3) as a fixed point iteration and considering the spectral radius of the fixed point iteration matrix.

Our new convergence analysis is tailored to the particular situation of having a Rayleigh functional, and differs significantly from [2, 3]. Our major motivation for reconsidering this question was to establish mesh-independent convergence rates when applying (3) with a multigrid preconditioner to the finite element discretization of a nonlinear PDE eigenvalue problem. The analyses in [2, 3] do not seem to admit such a conclusion, at least it is not obvious. On the other hand, such results are well known for the linear case $T(\lambda) = \lambda I - A$, for which (3) comes down to the preconditioned inverse iteration (PINVIT). In particular, the seminal work by Neymeyr establishes tight expressions for the convergence of the eigenvalue and eigenvector approximations produced by PINVIT. Unfortunately,

the elegance of Neymeyr's mini-dimensional analysis of the Rayleigh-quotient is strongly tied to linear eigenvalue problems; there seems little hope to carry it over to the general nonlinear case. Our approach proceeds by directly analysing the convergence of the eigenvector. Although leading to weaker bounds than Neymeyr's analysis in the linear case, the obtained results still allow to establish mesh-independent convergence rates.

In the first step, we show that

$$\tan \phi_P(v_{k+1}, x_1) \leq \gamma \cdot \tan \phi_P(v_k, x_1) + O(\varepsilon^2),$$

where x_1 is an eigenvector belonging to λ_1 and ϕ_P denotes the angle in the geometry induced by P. In the second step, we show that $\gamma < 1$ (independent of h) for a multigrid preconditioner of $T(\sigma)$ with σ sufficiently close to λ_1.

References

[1] Effenberger, E., Kressner, D.: On the convergence of the residual inverse iteration for nonlinear eigenvalue problems admitting a Rayleigh functionals. In preparation, 2013.

[2] Jarlebring, E., Michiels, W.: Analyzing the convergence factor of residual inverse iteration. BIT **51**(4), 937–957 (2011).

[3] Neumaier, A.: Residual inverse iteration for the nonlinear eigenvalue problem. SIAM J. Numer. Anal. **22**(5), 914–923 (1985)

[4] Voss, H.: Nonlinear eigenvalue problems. In: L. Hogben (ed.) Handbook of Linear Algebra. Chapman & Hall/CRC, FL (2013).

Preconditioning for Inexact Inner-Outer Methods for the Two-sided, Non-Hermitian Eigenvalue Problem

Melina Freitag and Patrick Kürschner

Abstract

We discuss the numerical solution of large-scale, two-sided, non-Hermitian eigenvalue problems

$$Ax = \lambda Mx \quad \text{and} \quad y^H A = \lambda y^H M$$

by iterative methods such as two-sided inverse, and Rayleigh quotient iteration [3] as well as two-sided Jacobi-Davidson [4]. The left eigenvectors y are required in some applications, e.g. eigenvalue and eigenvector based model order reduction of large dynamical systems. Their incorporation can also be beneficial for the performance of eigenvalue methods for non-normal problems.

In these two-sided eigenvalue iterations, adjoint linear systems of the form

$$(A - \theta_k M)u_{k+1} = Mu_k \quad \text{and} \quad (A - \theta_k M)^H v_{k+1} = M^H v_k \tag{1}$$

have to be solved in each iteration, where it is often sufficient to solve them inexactly, e.g. by employing Krylov subspace methods for linear systems. The iterations of this linear solver are commonly referred to as *inner iterations* in contrast to the *outer iterations* of the method for the eigenvalue computation.

In this presentation we focus on specially tailored preconditioning strategies for these inexact inner solves. It is known that [2], when a preconditioner $\mathbb{P}$ satisfies, e.g. $\mathbb{P}u_k = Mu_k$ for the first linear system in (1), the number of necessary inner iterations to obtain an approximation to u_{k+1} of a certain accuracy is significantly reduced. This leads to the concept of so called *tuned preconditioners*. A special property of (1) is that the coefficient matrices are adjoint to each other and hence, both linear systems can be dealt with simultaneously by iterative methods such as BiCG and QMR, but also GLSQR and the unsymmetric MINRES. The tuned preconditioners are adapted and tailored to the work in these inner solvers [1]. It turns out that this allows to draw novel connections between the aforementioned eigenvalue iterations.

If time permits some perspectives for tuned preconditioners for similar and related inner-outer methods, such as solvers for nonlinear eigenvalue problems and the dominant pole algorithm [5], are discussed. For instance, the dominant pole algorithms computes eigenvectors corresponding to certain eigenvalues for carrying out model order reduction of large dynamical systems. It has the special property that the right hand sides of the adjoint linear systems do not vary in the course of the iteration. This introduces additional complications for inexact solves.

References

[1] M. Freitag and P. Kürschner, *Tuned preconditioners for inexact two-sided inverse and Rayleigh quotient iteration*, MPI Magdeburg Preprint MPIMD Preprint/13-04, 2013. Available at `http://www.mpi-magdeburg.mpg.de/preprints/2013/04/`.

[2] M. Freitag, A. Spence, and E. Vainikko, *Rayleigh quotient iteration and simplified Jacobi-Davidson with preconditioned iterative solves for generalised eigenvalue problems*, technical report, Dept. of Mathematical Sciences, University of Bath, 2008.

[3] A. M. Ostrowski, *On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. III (generalized Rayleigh quotient and characteristic roots with linear elementary divisors)*, Archive for Rational Mechanics and Analysis, 3 (1959), pp. 325–240.

[4] M. E. Hochstenbach and G. L. G. Sleijpen, *Two-sided and alternating Jacobi–Davidson*, Linear Algebra and its Applications, 358(1-3) (2003), pp. 145–172.

[5] J. Rommes and G. L. G. Sleijpen, *Convergence of the dominant pole algorithm and Rayleigh quotient iteration*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 346–363.

Hierarchical QR Factorization Algorithms for Multi-Core Cluster Systems

Julien Langou

Abstract

In this presentation, we present our continuing research on tiled algorithms for QR factorization. Given an initial matrix, the tile QR factorization algorithm breaks down the matrix into tiles and then perform a QR factorization by performing sequential LAPACK-like kernels on the tiles [1,2,3]. Tile algorithms will naturally enables good data locality for the sequential kernels executed by the cores (high sequential performance), low number of messages in a parallel distributed setting (small latency term), and fine granularity (high parallelism). Each tile algorithm is uniquely characterized by its sequence of reduction trees. By changing the sequence of reduction trees, we can design algorithms with much less communication than LAPACK algorithms in two-level memory systems, algorithms with much less communication in parallel distributed systems than ScaLAPACK algorithms, algorithms with small granularity enabling large parallelism.

The general question is, given a set of resources, which data distribution and which sequence of trees are best to use. The goal is to maximize the parallelism of the factorization while minimizing the communication.

In essence, the tile algorithm is very similar to the Givens' rotation QR factorization algorithm; with the difference that, instead of operating on scalars (Givens' rotations), one operates on tiles. As far as Givens' rotations go, for square matrices, in 1978, Sameh and Kuck [4] gave an asymptotically optimal algorithm in term of shortest critical path length when no communications are considered. In 1986, Cosnard and Robert [5] gave an optimal algorithm for matrices of any shapes. In 2011, we [6] extended these results from Givens' rotations to tile algorithms. We proved that a Greedy schedule of the tasks related to the last column results in a shortest critical path length algorithm for matrices of any shapes. We demonstrated through experiments that this provides us with an efficient algorithm for current multicore platform.

We note that, in the multicore case, since our unit tasks operate on tiles, the ratio communication to computation per task has a surface to volume effect and so, one can neglect communications in a first approach design. The next step is to take into account communication, either in the multicore context where this is a plus, or, in the parallel distributed where this is a must.

This presentation [7] will describe a tile QR factorization algorithm which is especially designed for massively parallel platforms combining parallel distributed multi-core nodes. (These platforms represent the present and the foreseeable future of high-performance computing.) In the context of a cluster of multicores, in order to reduce the number of inter-processor communications, it is natural to consider hierarchical trees composed of an "inter-cluster" tree which acts on top of "intra-cluster" trees. At the intra-cluster level, we propose a hierarchical tree made of three levels: (0) "TS level" for cache-friendliness, (1) "low level" for decoupled highly parallel inter-node reductions, (2) "coupling level" to efficiently resolve interactions between local reductions and global reductions. Our hierarchical algorithm and its implementation are flexible and modular, and can accommodate several kernel types, different distribution layouts, and a variety of reduction trees at all levels, both inter-cluster and intra-cluster. Numerical experiments on a cluster of multicore nodes (i) confirm that each of the four levels of our hierarchical tree contributes to increase performance and (ii) build insights on how these levels influence performance and interact within each other. Our implementation of the new algorithm with the DAGuE scheduling tool [8] significantly outperforms currently available QR factorization softwares for all matrix shapes.

Beside experimental validation, we compare the performance model of our new algorithm to (1) ScaLAPACK algorithms and to (2) theoretical lower bounds. We show that our algorithm is asymptotically optimal in term of communication and computation and is much better than ScaLAPACK algorithms.

Bibliography

1. A. Buttari, J. Langou, J. Kurzak, and J. Dongarra. Parallel tiled QR factorization for multicore architectures. *Concurrency Computat.: Pract. Exper.*, 20(13):1573–1590, 2008.

2. A. Buttari, J. Langou, J. Kurzak, and J. Dongarra. A class of parallel tiled linear algebra algorithms for multicore architectures. *Parallel Computing*, 35:38–53, 2009.

3. G. Quintana-Ortí, E. S. Quintana-Ortí, R. A. van de Geijn, F. G. V. Zee, and E. Chan. Programming matrix algorithms-by-blocks for thread-level parallelism. *ACM Transactions on Mathematical Software*, 36(3):14:1–14:26, 2009.

4. A. Sameh and D. Kuck. On stable parallel linear systems solvers. *J. ACM*, 25:81-91, 1978.

5. M. Cosnard and Y. Robert. Complexity of parallel QR factorization. *J. ACM*, 33(4):712-723, 1986.

6. Henricus Bouwmeester, Mathias Jacquelin, Julien Langou, and Yves Robert. Tiled QR factorization algorithms. In *ACM/IEEE SC 2011 Conference (SC'11)*, November 2011.

7. Jack Dongarra, Mathieu Faverge, Thomas Herault, Mathias Jacquelin, Julien Langou, and Yves Robert. Hierarchical QR factorization algorithms for multi-core cluster systems. *Parallel Computing*, 39(4-5):212–232, 2013.

8. George Bosilca, Aurelien Bouteiller, Anthony Danalis, Thomas Herault, Pierre Lemarinier, Jack Dongarra. DAGuE: A generic distributed DAG engine for High Performance Computing. *Parallel Computing*, 38(1-2):37–51, 2012.

Hierarchical Preconditioners for Higher Order FEM

Sabine Le Borne

Abstract

The finite element discretization of a partial differential equation requires the selection of a suitable finite element space as well as a basis for this space. While higher order finite elements (HOFEMs) lead to finite element solutions of higher accuracy, their associated discrete linear systems of equations are often more difficult/costly to set up and to solve than those of lower order elements.

In this talk, we present efficient hierarchical preconditioners for the solution of linear systems of equations associated with HOFEMs. More specifically, we will develop "hybrid blackbox" preconditioners: The setup of the preconditioner will occur in a "blackbox" fashion, i.e., only the stiffness matrix is needed as input. However, the "hybrid" part implies that certain knowledge of the origin of the system is available and will possibly be exploited. Such knowledge could include a certain sparsity structure (e.g. produced through particular types of finite elements) or even a certain block structure (e.g. in mixed finite elements).

As a model problem, we consider the three-dimensional convection-diffusion problem with some non-constant, circulating convection. We use a regularly refined quadrilateral grid with the finite element space Q_p of continuous, piecewise polynomial elements of order p in each coordinate direction. As a basis, we will consider both a Lagrange (tensor) basis as well as a (p-) hierarchical basis. In either case, the resulting degrees of freedom can be classified into vertex, edge, face or interior (cell/bubble) nodes. While we restrict our attention here to this model problem, our approach and considerations in general are applicable to a much wider class of problems which will be pursued in future work. Here, we only make a note that we are not restricted to a certain type of model equation, structured grid or finite element space.

The preconditioners to be analysed and compared in this talk can be divided into the following four categories:

- approximate LU factorization of the entire matrix;
- block Gauß-Seidel method with approximate LU factorization of diagonal blocks;
- approximate LU factorization of the remaining matrix after static condensation of bubble dofs;
- block Gauß-Seidel method with approximate LU factorization of diagonal blocks after static condensation of bubble dofs.

The tool to be used for the efficient approximate LU factorizations required in all four categories is the technique of hierarchical ($\mathcal{H}$-) matrices. $\mathcal{H}$-matrices provide a powerful technique to compute and store approximations to dense matrices in a data-sparse format. The basic idea is the approximation of matrix data in hierarchically structured subblocks by low rank representations. Such low rank approximations can only be expected to be successful after suitable row and column permutations of the matrix. These permutations are typically determined in a blackbox fashion based on the underlying matrix graph of the (sparse) stiffness matrix.

In the past, $\mathcal{H}$-matrices have proven to be successful for the computation of approximate inverses as well as LU factors of stiffness matrices resulting from continuous, piecewise *linear* finite elements.

Here, we will analyze whether the $\mathcal{H}$-matrix technique is also successful in the context of higher order finite element discretizations.

As an introductory motivation, we turn to a simple one-dimensional model problem: the Laplace equation. Its discretization using p-th order finite elements leads to a stiffness matrix of bandwidth $2p+1$ that has a dense inverse. We will analyse $\mathcal{H}$-matrix representations of these exact inverses as well as their approximations.

Returning to our three-dimensional model convection-diffusion problem, we will consider $\mathcal{H}$-LU factorizations for decreasing grid width h (and fixed polynomial degree p) as well as for increasing polynomial degree p (and fixed h), both for varying convection dominance in our model problem.

An important aspect is the treatment of dense subblocks corresponding to the bubble functions associated with a finite element. Starting with a cube subdivided into eight subcubes, the final stiffness matrix will contain 8^ℓ dense blocks each of size $(p-1)^3 \times (p-1)^3$ (where ℓ denotes the number of refinements). Such subblocks can be identified in a blackbox manner even if the stiffness matrix is not (yet) sorted accordingly. While these blocks are unproblematic for smaller p, one needs to think about an efficient treatment of these blocks as p increases, say beyond $p = 6$. This holds particularly true in the case of static condensation of these blocks: While their elimination from the system does not lead to any additional fill-in, it can become quite costly and a bottleneck in the overall computational setup time. One possible remedy would be the approximation and subsequent elimination of these blocks in the $\mathcal{H}$-matrix format. The challenge here is the determination of a suitable row and column permutation prior to the $\mathcal{H}$-approximation since the underlying matrix graph is complete.

While the construction of an $\mathcal{H}$-LU approximation as a preconditioner is considered to be efficient if its work complexity is almost linear (up to logarithmic factors), it can still be costly in practice due to high constants involved in these complexity bounds. Thus, we also pursue block Jacobi and block Gauß-Seidel approaches in which only blocks on the diagonal are factorized instead of the entire matrix. As a drawback, we no longer obtain an (almost) exact LU-factorization - and therefore an exact/direct solver - as we increase the accuracy of the $\mathcal{H}$-arithmetic by allowing higher ranks in the subblocks. The benefit, however, are reduced setup times, and oftentimes these savings in setup times are much larger than the increase in the iteration times, leading to overall faster solution methods. The blocks to be used for such a block Jacobi or Gauß-Seidel method can be determined by the classification of the degrees of freedom through their association with vertices, edges, faces or cells, which is detectable from the sparsity structure of the matrix. Especially in the case of a hierarchical basis, such an approach promises to be successful since the linear basis functions are the ones associated with vertices, and the quadratic basis functions are included in those associated with edges. Basis functions of order higher than two, however, will be difficult to separate.

We will analyze this "trade-off" between setup times and iteration times both for increasing polynomial finite element degree p and decreasing grid width h.

All the introduced preconditioners will be illustrated with supporting numerical results. We will show setup times, storage requirements and iteration steps/times to reach an iterative solution within a desired accuracy.

Matrix Iterations and Pták's Method of Nondiscrete Induction

Jörg Liesen

Abstract

In the late 1960s, Vlastimil Pták (1925–1999) invented the *method of nondiscrete induction*, which is a method for estimating the convergence of iterative processes. Pták described the motivation for this method in his paper "What should be a rate of convergence?" [5] as follows:

> "It seems therefore reasonable to look for another method of estimating the convergence of iterative processes, one which would satisfy the following requirements.
>
> 1° it should relate quantities which may be measured or estimated during the actual process
>
> 2° it should describe accurately in particular the initial stage of the process, not only its asymptotic behaviour since, after all, we are interested in keeping the number of steps necessary to obtain a good estimate as low as possible."

This seems to be almost too much to ask for. Yet, for some iterations the method of nondiscrete induction indeed leads to analytical convergence estimates which satisfy both of Pták's requirements. To briefly describe the idea, suppose that a sequence of iterates x_k converges to the solution x_* of a given problem. Instead of a "classical" convergence bound of the form

$$\|x_* - x_{k+1}\| \leq c \, \|x_* - x_k\|^p,$$

where $c > 0$ is some constant and the *number* $p > 1$ is the "convergence rate" (e.g. "quadratic convergence" if $p = 2$), the method of nondiscrete induction gives easily computable bounds of the form

$$\|x_* - x_{k+1}\| \leq \sigma(\|x_{k+1} - x_k\|),$$

where σ is a *scalar function*. For a few examples, including Newton's method [3] and an iteration for solving a certain algebraic eigenvalue problem [4], Pták derived such a function σ by comparing consecutive increments of the iteration. His joint monograph with Potra [2] summarizes the most important results.

A footnote in [4] says that Pták "intended" to present the method of nondiscrete induction at the *Symposium on Numerical Algebra, Gatlinburg VI*, held in December 1974 in Hopfen am See, Bavaria, Germany. Later Pták referred to [4] as his "Gatlinburg Lecture". According to some participants, however, Pták did not attend Gatlinburg VI and thus he missed that particular chance to present his method to the numerical linear algebra community. Apparently, the method of nondiscrete induction did not become widely known in this community, and even in the literature on Newton methods for nonlinear problems the method is often mentioned only marginally, if at all. Part of the reason for this neglect of the method may be the rather theoretical nature of the original publications, which are formulated in general (Banach space) settings, as well as their lack of numerical examples.

Based on [1], the goals of this talk are, first, to explain on an example how the method of nondiscrete induction can be applied to a matrix iteration in numerical linear algebra. For this purpose we will use the Newton iteration for the matrix square root, and possibly present other applications, if time allows. Second, it will be illustrated numerically that the method can be quite effective, at least in some cases. The talk will thus reveal the essential ideas of the method, the main requirements for its applicability in the context of matrix iterations, and its strengths and weaknesses.

References

[1] J. Liesen, *Pták's nondiscrete induction and its application to matrix iterations*, arXiv:1405.2683, submitted, 2014.

[2] F. A. Potra and V. Pták, *Nondiscrete Induction and Iterative Processes*, vol. 103 of Research Notes in Mathematics, Pitman (Advanced Publishing Program), Boston, MA, 1984.

[3] V. Pták, *The rate of convergence of Newton's process*, Numer. Math., 25 (1975/76), pp. 279–285.

[4] ——, *Nondiscrete mathematical induction and iterative existence proofs*, Linear Algebra and Appl., 13 (1976), pp. 223–238.

[5] ——, *What should be a rate of convergence?*, RAIRO Anal. Numér., 11 (1977), pp. 279–286.

Covariance Structure Regularization via Entropy Loss Function

Lijing Lin, Nicholas J. Higham and Jianxin Pan

Abstract

The need to estimate structured covariance matrices arises in a variety of applications such as signal processing [5], networks [6], and automatic control [4], and the problem is widely studied in statistics. A conventional way, known as the "Burg technique", is to find the maximum likelihood estimation for a covariance matrix that has a specific/regularized structure using random samples drawn from a stochastic process [1]. However, this method has some drawbacks: (a) it is based on the presumption that the stochastic process is multivariate normal; (b) the structure of the covariance must be prespecified; (c) the sample covariance matrix must be available; moreover, deducing the underlying covariance structure from the sample covariance matrix can be difficult due to random noise and the large dimension of that matrix.

To overcome these difficulties, we propose a new method for regularizing the underlying structure of a given covariance matrix. Our method is based on the *entropy loss function* [2]

$$L(A,B) = \operatorname{trace}(A^{-1}B) - \log(\det(A^{-1}B)) - m, \tag{1}$$

defined on two matrices A and B of size $m \times m$ and, to ensure that $L(A,B)$ is nonnegative, we assume that A and B are symmetric positive definite. The entropy loss function, also known as the Kullback-Leibler divergence, is a well-accepted nonsymmetric measure of the discrepancy between two probability distributions. The problem of interest here is, given a covariance matrix A whose underlying structure is blurred due to random noise, particularly when the dimension m is high, to identify the underlying structure of A from a class of candidate covariance structures.

Assume that we have a class of k candidate structures $\{s_1, s_2, \ldots, s_k\}$ and that the given covariance matrix A is positive definite. Let $\mathcal{S}_i$ be the set of all positive definite covariance matrices with structure s_i. We define the discrepancy between A and $\mathcal{S}_i$ as

$$D(A, \mathcal{S}_i) = \min_{B \in \mathcal{S}_i} L(A,B), \tag{2}$$

where $L(A,B)$ is the entropy loss function in (1). The idea is to find the set $\mathcal{S}_* \in \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ that is "closest" to A:

$$\mathcal{S}_* = \arg\min_{1 \le i \le k} D(A, \mathcal{S}_i),$$

and take the corresponding structure of $\mathcal{S}_*$ as the most likely underlying structure of A. We then refer to the replacement of A by a matrix $B \in \mathcal{S}_*$ achieving the minimum in (2) as the process of regularization.

In this talk we consider a range of four candidate covariance structures that are commonly used in practice, for example, in longitudinal and spatial studies.

(1) The order-1 moving average structure, **MA(1)**, has a tridiagonal Toeplitz covariance matrix

$$B = \sigma^2 \begin{bmatrix} 1 & c & \cdots & 0 \\ c & 1 & \ddots & \vdots \\ \vdots & \ddots & 1 & c \\ 0 & \cdots & c & 1 \end{bmatrix}.$$

(2) The covariance of compound symmetry (**CS**) structure assumes the same correlation coefficient for every pair of observations:

$$B = \sigma^2 \begin{bmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \ddots & \ddots & \vdots \\ c & \cdots & c & 1 \end{bmatrix}.$$

(3) Autoregression of order 1, **AR(1)**, has a covariance where the correlation between any pair of observations decays exponentially towards zero as the distance between them increases:

$$B = \sigma^2 \begin{bmatrix} 1 & c & c^2 & \cdots & c^{m-1} \\ c & 1 & c & \cdots & c^{m-2} \\ c^2 & c & 1 & \cdots & c^{m-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c^{m-1} & c^{m-2} & \cdots & c & 1 \end{bmatrix}.$$

(4) More generally, **banded Toeplitz** covariance matrices have constant subdiagonals, i.e., constants at lag 1 , lag 2, . . . , and lag p:

$$B = \sigma^2 \begin{bmatrix} 1 & c_1 & \cdots & c_p & \cdots & 0 \\ c_1 & 1 & c_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & c_p \\ c_p & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & c_1 \\ 0 & \cdots & c_p & \cdots & c_1 & 1 \end{bmatrix}.$$

The main task now is to calculate the discrepancy $D(A, \mathcal{S})$ for each of the candidate covariance structures: MA(1), CS, AR(1), and banded Toeplitz. We show that for the first three structures local or global minimizers of the discrepancy can be computed by one-dimensional optimization, while for the fourth structure Newton's method enables efficient computation of the global minimizer. Simulation studies are conducted, showing that the proposed new approach provides a reliable way to regularize covariance structures. The approach is also applied to real data analysis, demonstrating the usefulness of the proposed approach in practice.

This talk is based on [3].

References

[1] J. P. Burg, D. G. Luenberger, and D. L. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, 1982.

[2] W. James and C. Stein. Estimation with quadratic loss. In J. Neyman, editor, *Proceedings of the Fourth Berkeley Symposium*, volume 1 of *Mathematical Statistics and Probability*, pages 361–379. University of California Press, 1961. The Statistical Laboratory, University of California, June 30-July 30, 1960.

[3] Lijing Lin, Nicholas J. Higham, and Jianxin Pan. Covariance structure regularization via entropy loss function. MIMS EPrint 2012.61, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, June 2012. Revised October 2013. To appear in Comput. Statist. Data Anal.

[4] F. Lin and M. R. Jovanović. Least-squares approximation of structured covariances. *IEEE Trans. Automat. Control*, 54(7):1643–1648, 2009.

[5] F. Pascal, Y. Chitour, J. P. Ovarlez, P. Forster, and P. Larzabal. Covariance structure maximum-likelihood estimates in compound Gaussian noise: existence and algorithm analysis. *IEEE Trans. Signal Processing*, 56(1):34–48, 2008.

[6] V. Vinciotti and H. Hashem. Robust methods for inferring sparse network structures. *Computational Statistics & Data Analysis*, 67:84–94, 2013.

The Hyperbolic Quadratic Eigenvalue Problem

Ren-Cang Li and Xin Liang

Abstract

The hyperbolic quadratic eigenvalue problem (HQEP) was shown to admit the Courant-Fischer type min-max principles in 1955 by Duffin [1] and Cauchy type interlacing inequalities in 2010 by Veselić [4]. It can be regarded as the closest analogue (among all kinds of quadratic eigenvalue problems) of the standard Hermitian eigenvalue problem (among all kinds of standard eigenvalue problems) [2]. In this talk, we will present our recent study on HQEP both theoretically and numerically [3]. In the theoretic front, we generalize Wiedlandt-Lidskii type min-max principles and, as a special case, Ky Fan type trace min/max principles and establish Weyl type and Mirsky type perturbation results when an HQEP is perturbed to another HQEP. In the numerical front, we justify the natural generalization of the Rayleigh-Ritz procedure with the existing and our new optimization principles and, as consequences of these principles, we extend various current optimization approaches – steepest descent/ascent and nonlinear conjugate gradient type methods for the Hermitian eigenvalue problem – to calculate few extreme quadratic eigenvalues (of both pos- and neg-type). A detailed convergent analysis is given on the steepest descent/ascent methods. The analysis reveals the intrinsic quantities that control convergence rates and consequently yields ways of constructing effective preconditioners. Numerical examples are presented to demonstrate the proposed theory and algorithms.

References

[1] R. Duffin. A minimax theory for overdamped networks. *Indiana Univ. Math. J.*, 4:221–233, 1955.

[2] Nicholas J. Higham, Françoise Tisseur, and Paul M. Van Dooren. Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems. *Linear Algebra Appl.*, 351-352:455–474, 2002.

[3] Xin Liang and Ren-Cang Li. The hyperbolic quadratic eigenvalue problem. work-in-progress (80 pages), 2013.

[4] K. Veselić. Note on interlacing for hyperbolic quadratic pencils. In Jussi Behrndt, Karl-Heinz Förster, and Carsten Trunk, editors, *Recent Advances in Operator Theory in Hilbert and Krein Spaces*, volume 198 of *Oper. Theory: Adv. Appl.*, pages 305–307. 2010.

Improved Divide-and-Conquer Algorithms for the Eigenvalue and Singular Value Problems

Shengguo Li, Ming Gu, Lizhi Cheng and Xuebin Chi

Abstract

Efficient and reliable rank-structured matrix computations have been an intensive focus of recent research. Rank-structured matrices includes semiseparable, quasiseparable, sequentially semiseparable(SSS), hierarchically semiseparable(HSS), $\mathcal{H}$, $\mathcal{H}^2$ matrices, and etc. They have been used for solving Toeplitz linear equations, integral equations, computing the zeros of polynomials, developing structured sparse solvers and so on. In this talk, we present an algorithm of using HSS matrices to accelerate the computation of bidiagonal SVD and tridiagoanl eigenvalue problems. Since these two problems are quite related, we mainly talk about the bidiagonal case.

The main cost of divide-and-conquer (DC) algorithm for the bidiagonal SVD problems lies to the computation of singular vectors, which needs two matrix-matrix multiplications. Recall that the DC algorithm [2] needs to compute the SVD of a sequences of *broken arrow* matrices. An important fact we found is that the singular vector matrices of a broken arrow matrix are *Cauchy-like matrices* and have *off-diagonal low-rank property*. To exploit these two properties, we first use a structured low-rank approximation method to construct an HSS approximation to each Cauchy-like matrix, which costs $O(K^2r)$ flops with $O(K)$ memory, where K is the size of matrix and r is the maximum rank of its off-diagonal blocks, which is a medium constant, say around 20-50. Then the singular vectors are updated via fast HSS matrix-matrix multiplications, which also costs $O(K^2r)$ flops [3].

The approximation error of the HSS construction algorithm is analyzed, when uses the structured low-rank approximation algorithm for Cauchy-like matrices. That structured approximation algorithm is proved to be backward stable. To further show its stability, we tested it by using the matrices in LAPACK tester `stetester` [4]. The implementation details are included. Numerous experiments are done on a laptop with 4GB memory and Intel(R) Core(TM) i7-2640M CPU, and the codes are compiled by Intel fortran compiler (`ifort`) with optimization flag `-O2`. When comparing with highly optimized BLAS libraries such as Intel MK, our algorithm can be more than 3x faster for some large matrices with few deflation. Since HSS construction [1] and HSS multiplication algorithms [3] are naturally parallelizable, we further implement these two algorithms in parallel by using OpenMP. The numerical results show that we can get good speedups.

References

[1] S. Chandrasekaran, M. Gu and T. Pals, *A fast ULV decomposition solver for hierarchically semiseparable representations*, SIAM J. Matrix Anal. Appl., 28(2006): 603–622.

[2] M. Gu and S.C. Eisenstat *A divide-and-conquer algorithm for the bidiagonal SVD*, SIAM J. Matrix Anal. Appl., 19(1998): 79–92.

[3] W. Lyons *Fast algorithms with applications to PDEs*, PhD thesis, University of California, Santa Barbara, 2005.

[4] Q.A. Marques, C. Voemel, J.W. Demmel and B.N. Parlett, *Algorithm 880: A testing infrastructure for symmetric tridiagonal eigensolvers*, ACM Trans. Math. Softw., 35(2008).

Hierarchically Low-Rank Structured Sparse Factorization with Reduced Communication and Synchronization

Pieter Ghysels, Xiaoye S. Li, Artem Napov, François-Henry Rouet and Jianlin Xia

Abstract

We have been developing a new class of sparse LU factorization algorithms and codes with low arithmetic and communication complexity, which are targeted at broad classes of PDEs at extreme scale, taking advantage of data sparseness in the discretized matrices. It was long discovered that for structured matrices, such as Hilbert matrix, Toeplitz matrix and the matrices from the BEM methods for integral equations, the off-diagonal blocks are numerically rank deficient. Several compression techniques using hierarchically truncated SVD representation have been develoeped. These compact forms can be used to design asymptotically faster and low-memory linear algebra algorithms [1, 3]. In recent years, these structured matrix ideas were introduced in the sparse solvers for solving discretized PDEs [2, 10]. In particular, we have pioneered the work of using *hierarchically semi-separable* (HSS) structure to develop scalable superfast sparse direct solver. We developed parallel algorithms for the key HSS operations, including construction, ULV factorization and solution [8]. Moreover, we integrated these parallel HSS kernels into a parallel geometric sparse multifrontal method, and demonstrated that our HSS-embedded sparse direct solver can solve the discretized Helmholtz equations with 6.4 billion unknowns using 16,000+ processing cores [7]. This is beyond reach of the traditional direct solvers as well as many other types of solvers.

Despite the initial success, we encountered one major road block: the compression kernel based on the traditional rank-revealing QR (RRQR) method is difficult to scale up. Worse yet, it leads to non-compatible HSS structures from different branches of the elimination tree, which prevents fast assembling of Schur complement updates at each step of Gaussian elimination (i.e., *extend-add* of the children's update matrices into the parent frontal matrix.) We found that the *randomized sampling* (RS) method can replace RRQR to obtain the desired truncated SVD compression. The following procedure computes an orthonormal basis (ON-basis) U for an $m \times n$ matrix B such that $\|B - UU^*B\|$ is small *with high probability.*

1. Choose a Gaussian random matrix $\Omega_{n\times(r+p)}$, where r is the expected rank and p is a small over-sampling constant of $O(1)$.
2. Form the sample matrix $S = B\,\Omega$ whose columns span the column space of B.
3. Construct an ON-basis U for S using a strong RRQR.

It can be shown that U is a good ON-basis for B's column space [4]. The major saving comes from the fact that the column dimension of the sampling matrix S is significantly smaller. This procedure can be used in each step of the HSS-embedded multifrontal factorization as follows: 1) using the RS procedure to construt HSS forms for the frontal and update matrices [5]; 2) form a sample update matrix by multiplying the update matrix with a random matrix; 3) perform extend-add of the *sampled* update matrix to the parent. The advantages are:

- Extend-add involves tall-skinny dense sample matrices of the same shape, instead of the non-compatible HSS structures.
- The main kernel operation is dense matrix-matrix multiplication, which is well studied for good scalability and resilience at large scale.
- It is especially attractive when the matrix from the application admits FMM structure which leads to very fast matrix-vector product, or when the application cannot afford to store the matrix but only has matrix-vector operator; then we can develop the *matrix-free* sparse factorization algorithms.

This randomized sparse factorization framework can be used to construct nearly linear-time sparse direct solver and preconditioner [9], sparse eigensolver, selected inversion, among others. Although this type of solver/preconditioner is not as black-box as a direct solver, it is applicable to broad classes of PDEs, including some of those for which multigrid does not work [6].

In this talk, we will present the following new results: 1) New adaptive algorithm to accommodate rank variations at different parts of the matrix: 2) New scalable HSS-embedded sparse factorization method that employs hierarchical parallelism both at the algorithm level and at the architectural level (e.g., multicore/manycore-awareness); 3) Analysis and demonstration of reduced memory access and communication complexity of the new algorithm compared to the traditional factorization methods; and 4) Parallel performance of the algorithm on modern HPC machines, when it is used both as a direct solver and as a preconditioner.

References

[1] M. Bebendorf. *Hierarchical Matrices*, volume 63 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin Heidelberg, 2008.

[2] B. Engquist and L. Ying. Sweeping preconditioner for the helmholtz equation: hierarchical matrix representation. *Commun. Pure Appl. Math.*, 64:697–735, 2011.

[3] W. Hackbusch, L. Grasedyck, and S. Börm. An introduction to hierarchical matrices. *Math. Bohem.*, 127:229–241, 2002.

[4] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[5] P.G. Martinsson. A fast randomized algorithm for computing a hierarchically semiseparable representation of a matrix. *SIAM J. Matrix Analysis and Applications*, 32(4):1251–1274, 2011.

[6] A. Napov, X.S. Li, and M. Gu. An algebraic multifrontal preconditioner that exploits a low-rank property. Technical report, Lawrence Berkeley National Laboratory, 2013. In preparation.

[7] S. Wang, X.S. Li, F.-H. Rouet, J. Xia, and M.V. de Hoop. A parallel fast geometric multifrontal solver using hierarchically semiseparable structure. *ACM Trans. Mathematical Software*, 2013. (submitted).

[8] S. Wang, X.S. Li, J. Xia, Y. Situ, and M.V. de Hoop. Efficient parallel algorithms for solving linear systems with hierarchically semiseparable structures. *SIAM J. Scientific Computing*, 2013. (to appear).

[9] J. Xia. Randomized sparse direct solvers. *SIAM J. Matrix Analysis and Applications*, 34:197–227, 2013.

[10] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li. Superfast multifrontal method for large structured linear systems of equations. *SIAM J. Matrix Anal. Appl.*, 31:1382–1411, 2009.

Quasi-Canonical Forms for Quadratic Matrix Polynomials

D. Steven Mackey, F. De Terán, F. Dopico, Vasilije Perović and Francoise Tisseur

Abstract

The Weierstrass and Kronecker canonical forms for matrix pencils [2] are indispensable tools for obtaining insight into both the theoretical and computational behavior of pencils and their corresponding eigenproblems. The absence of any analogous result for matrix polynomials of higher degree has made it difficult to achieve the same depth of understanding for general matrix polynomials as we have for pencils. In this talk I will describe recent progress towards canonical forms for *quadratic* matrix polynomials, both for general quadratics (square and rectangular) as well as for polynomials in various important structure classes such as Hermitian and palindromic quadratic matrix polynomials.

As a first step towards resolving the quadratic canonical form question, it is convenient to consider a given list $\mathcal{L}$ of elementary divisors (finite and/or infinite) together with (left and/or right) minimal indices as the initial data. Then the *quadratic realizability problem* (QRP) consists of two basic issues:

- *Characterize* those lists $\mathcal{L}$ that comprise the complete spectral and singular structure of some quadratic matrix polynomial.
- For each such list $\mathcal{L}$, show how to *concretely construct* a quadratic matrix polynomial that realizes the list $\mathcal{L}$. It is also desirable for this concrete realization to be as simple and canonical as possible.

Structured versions of the QRP can also be formulated — given a class $\mathcal{S}$ of structured matrix polynomials, characterize (and constructively realize) those lists $\mathcal{L}$ that comprise the complete spectral and singular structure of some quadratic matrix polynomial *in* $\mathcal{S}$. Thus we have the palindromic QRP, the alternating QRP, the Hermitian QRP, ..., etc.

Solutions to many of these QRPs have recently been obtained by developing *quasi-canonical forms* for both structured and unstructured quadratic matrix polynomials. These quasi-canonical forms are direct sums of canonical Kronecker-like quadratic blocks, each of which realizes a "minimal" list of elementary divisors and minimal indices.

The development of these quasi-canonical forms has revealed several new phenomena that occur for quadratic polynomials, but not for matrix pencils.

(a) Although *any* list $\mathcal{L}$ can be realized by a matrix pencil (simply build a Kronecker block for each individual elementary divisor and for each minimal index), this is not true in the quadratic case. There are necessary conditions on $\mathcal{L}$ for quadratic realizability arising from the Index Sum Theorem [1]; for structured quadratic realizability there are additional necessary conditions arising from the restricted Smith forms of these structure classes [3, 4, 5].

(b) There exist nontrivial *quadratically irreducible* lists $\mathcal{L}$, i.e., quadratically realizable lists $\mathcal{L}$ with at least two elements that *cannot be partitioned* into quadratically realizable sublists. Clearly each quadratically irreducible list must be realized by its own individual block; the existence of such lists thus constitutes an important reason why the QRP is significantly more complicated than the corresponding realizability problem for pencils. Identifying all of the qualitatively distinct types of quadratically irreducible lists is one of the main contributions of this work.

(c) There exist nontrivial quadratically irreducible lists $\mathcal{L}$ that contain a mixture of both elementary divisors *and* minimal indices. The existence of such lists shows that it is not always possible to cleanly separate a quadratic matrix polynomial into a "regular part" and a "singular part", as can always be done for pencils.

(d) The canonical blocks in the classical Weierstrass and Kronecker forms for pencils are all bidiagonal. In the quadratic case, though, it can be shown that *bidiagonal blocks do not suffice*; i.e., there are quadratically irreducible lists $\mathcal{L}$ that are not realizable by any bidiagonal quadratic matrix polynomial. Thus the canonical blocks in any quadratic canonical form must necessarily be more complicated than those in the canonical forms for pencils.

The condensed quadratic realizations developed in this work are built as direct sums of quadratic blocks, each of which is canonical for some quadratically irreducible sublist of the given list $\mathcal{L}$ of elementary divisors and minimal indices. However, the quadratic condensed form as a whole is *not always unique*; there exist quadratically realizable lists $\mathcal{L}$ that can be partitioned into quadratically irreducible sublists in more than one way, leading to qualitatively distinct quadratic realizations. For this reason (and others to be discussed in the talk), the quadratic condensed forms presented here cannot be regarded as canonical forms in the usual sense of the term. Hence we refer to these condensed forms only as *quasi*-canonical forms, since they possess only some of the properties associated with a true canonical form.

References

[1] F. De Terán, F. Dopico, and D. S. Mackey. Spectral equivalence of matrix polynomials and the index sum theorem. Available as MIMS EPrint 2013.47, Manchester Institute for Mathematical Sciences, Manchester, England, 2013.

[2] F. R. Gantmacher. *Theory of Matrices*. Chelsea, New York, 1959.

[3] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Jordan structures of alternating matrix polynomials. *Linear Algebra Appl.*, 432(4):867–891, 2010.

[4] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Smith forms of palindromic matrix polynomials. *Electron. J. Linear Algebra*, 22:53–91, 2011.

[5] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Skew-symmetric matrix polynomials and their Smith forms. *Linear Algebra Appl.*, 438(12):4625–4653, 2013.

Recent Results in Randomized Numerical Linear Algebra

Michael W. Mahoney

Abstract

Matrix algorithms and numerical linear algebra provide the foundation for many methods in scientific computing, engineering, machine learning, and data analysis, and in recent years randomization has proven to be a powerful if unexpected resource in the development of qualitatively improved matrix algorithms that are designed to address increasingly-common large-scale statistical data analysis problems. Here, we will review several recent developments in this area of Randomized Numerical Linear Algebra: first, an algorithm to compute a low-precision solution to an arbitrary over-constrained least-squares problem in time that is proportional to the number of nonzero elements in the input, plus lower order terms, as well as extensions of this to low-rank and leverage score approximation; second, recent improvements in Nystrom-based approximation of symmetric positive semidefinite matrices; and third, a framework to begin to address the statistical properties of random sampling and random projection algorithms. In each case, we will see that these developments depend crucially on understanding and exploiting the complementary algorithmic and statistical properties of the empirical statistical leverage scores of the input matrices. Extensions of the ideas underlying leverage-based randomized algorithms to address problems beyond the traditional purview of Randomized Numerical Linear Algebra may also be briefly described.

Structured Low-Rank Approximation with Missing Data

Ivan Markovsky and Konstantin Usevich

Abstract

We consider the problem of approximating an affinely structured matrix with missing elements by a low-rank matrix with the same structure. The method proposed is based on reformulation of the problem as inner and outer minimization. The inner minimization is a singular linear least-norm problem with analytic solution. The outer problem is a nonlinear least squares problem and is solved by local optimization methods. The method is generalized to weighted low-rank approximation with missing values and is illustrated on approximate low-rank matrix completion, system identification, and model-free control problems.

Problem formulation Given a data vector $p \in (\mathbb{R} \cup \{\,\texttt{NaN}\,\})^{n_p}$ (`NaN` denotes missing values), matrix structure—an affine function $\mathcal{S} : \mathbb{R}^{n_p} \to \mathbb{R}^{m \times n}$ from the structure parameter space $\mathbb{R}^{n_p}$ to the set of matrices $\mathbb{R}^{m \times n}$, and a bound $r < m \leq n$ on the rank

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{p} \in \mathbb{R}^{n_p} \quad \sum_{\{\,i \;\mid\; p_i \neq \texttt{NaN}\,\}} (p_i - \widehat{p}_i)^2 \\ \text{subject to} & \operatorname{rank}\big(\mathcal{S}(\widehat{p})\big) \leq r. \end{array} \tag{SLRA}$$

Solution approach Using the kernel representation of the rank constraint

$$\operatorname{rank}\big(\mathcal{S}(\widehat{p})\big) \leq r \quad \Longleftrightarrow \quad \text{there is } R \in \mathbb{R}^{(m-r)\times m}, \text{ such that } R\mathcal{S}(\widehat{p}) = 0 \text{ and } R \text{ has full row rank,}$$

the following equivalent problem to (SLRA) is obtained:

$$\begin{array}{ll} \text{minimize} & \text{over } R \in \mathbb{R}^{(m-r)\times m} \quad M(R) \\ \text{subject to} & R \text{ has full row rank,} \end{array} \tag{OUTER}$$

where

$$M(R) := \min_{\widehat{p} \in \mathbb{R}^{n_p}} \sum_{\{\,i \;\mid\; p_i \neq \texttt{NaN}\,\}} (p_i - \widehat{p}_i)^2 \quad \text{subject to} \quad R\mathcal{S}(\widehat{p}) = 0. \tag{INNER}$$

The evaluation of the cost function M of (OUTER), *i.e.*, solving (INNER) for a given value of R, is refered to as the inner minimization problem. This problem is a generalized least norm problem and is solved analytically (see Lemma 1). The problem of minimizing M over R is refered to as the outer minimization problem. It is a nonlinear least-squares problem with a full rank constraint of the parameter matrix R, *i.e.*, an *optimization problem on a Stiefel manifold* [1]. The full rank constraint makes the method proposed for solving (SLRA) a non-standard *variable projection method* [2].

Lemma 1 (Generalized least norm problem)**.** *Consider the generalized linear least norm problem*

$$\begin{array}{ll} \textit{minimize} & \textit{over } x \in \mathbb{R}^{n_x} \textit{ and } y \in \mathbb{R}^{n_y} \quad \|x\|_2^2 \\ \textit{subject to} & Ax + By = c, \end{array} \tag{GLN}$$

with $A \in \mathbb{R}^{m \times n_x}$, $B \in \mathbb{R}^{m \times n_y}$, and $c \in \mathbb{R}^m$. Under the following assumptions:

1. *B is full column rank,*
2. *$1 \leq m - n_y \leq n_x$, and*

3. $\bar{A} := B^{\perp}A$ *is full row rank,*

problem (GLN) has a unique solution

$$x = \bar{A}^{\top}\big(\bar{A}\bar{A}^{\top}\big)^{-1}B^{\perp}c \qquad \textit{and} \qquad y = B^{+}(c - Ax).$$

The minimum is

$$f = c^{\top}(B^{\perp})^{\top}\big(\bar{A}\bar{A}^{\top}\big)^{-1}B^{\perp}c.$$

There is a link between the generalized least norm problem (GLN) and the weighted least norm problems

$$\min_{z} \; z^{\top}Wz \quad \text{subject to} \quad Dz = c,$$

with a singular weight matrix W.

Applications In the case of unstructured matrix, (SLRA) is an approximate matrix completion problems. In the case of Hankel structured matrix, (SLRA) is a system identification problem for data with missing input and output measurements. The inner minimization problem (INNER) is in this case a Kalman smoothing problem. Efficient $O(n)$ computation is possible by exploiting the matrix structure. The algorithm can be viewed as a generalization of the classical Kalman smoother for the case of missing data.

Another application of (SLRA) is model-free control, *i.e.*, computation of a (feed-foreword) control signal for a dynamical system that is specified implicitly by a noisy trajectory. The classical approach to solve the control problem is to identify a model of the system from the data (system identification problem) and use the model in a second step (model-based control problem) for the construction of the control signal. The model-free (SLRA) approach computes the control signal directly from data, bypassing the explicit model identification step.

More information A paper [3] and software for structured low-rank approximation with missing data are available from `http://slra.github.io`

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Institute of Physics, Inverse Problems*, 19:1–26, 2003.

[3] I. Markovsky and K. Usevich. Structured low-rank approximation with missing data. *SIAM J. Matrix Anal. Appl.*, pages 814–830, 2013.

[4] C. Paige. Fast numerically stable computations for generalized linear least squares problems. *SIAM J. Numer. Anal.*, 16:165–171, 1979.

On Solving KKT Linear Systems arising in Model Predictive Control via Recursive Anti–Triangular Factorization

Nicola Mastronardi, Paul Van Dooren and Raf Vandebril

Abstract

The solution of Model Predictive Control problems [2, 3, 4] is often computed in an iterative fashion, requiring to solve, at each iteration, a quadratic optimization problem. The most expensive part of the latter problem is the solution of symmetric indefinite KKT systems, where the involved matrices are highly structured.

Recently, an algorithm for computing a block anti–triangular factorization of symmetric indefinite matrices, based on orthogonal transformations, has been proposed [1]. The aim of this talk is to show that such a factorization, implemented in a recursive way, can be efficiently used for solving the above mentioned KKT linear systems.

References

[1] N. Mastronardi, P. Van Dooren, *The anti–triangular factorization of symmetric matrices*, SIAM Journal on Matrix Analysis and Applications, 34(1), pp. 173-196, 2013.

[2] C. Kirches, H. Bock, J.P. Schlöder, S. Sager, *A factorization with update procedures for a KKT matrix arising in direct optimal control*, Mathematical Programming Computation, 3(4), pp. 319-348, 2011.

[3] Y. Wang, S. Boyd, *Fast Model Predictive Control Using Online Optimization*, IEEE Transactions on Control Systems Technology, 18(2), pp. 267-278, 2010.

[4] V.M. Zavala, C.D. Laird, L,T. Biegler, *A fast moving horizon estimation algorithm based on nonlinear programming sensitivity*, Journal of Process Control, 18(9), pp. 876–884, 2008.

Tensor Padé Krylov Methods for Parametric Model Order Reduction

Karl Meerbergen

Abstract

Consider the linear parametric single-input single-output system

$$\begin{aligned} \frac{dx(t,\gamma)}{dt} + A(\gamma)x(t,\gamma) &= fu(t) \quad , \quad x(0,\gamma) \equiv 0 \\ y(t,\gamma) &= c^* x(t) \end{aligned} \tag{1}$$

where x, f and c are n-vectors with n large, A is large and sparse and depends on parameters $\gamma \in \mathbb{R}^p$. Such problems arise in parametric studies of, e.g., electrical circuits, acoustics and vibrations [1]. When γ is a vector of stochastic parameters, we may be interested in the stochastic moments of the output y. The mean (which is the 0-th moment), e.g., is

$$z(t) = \int_\Gamma y(t,\gamma) d\gamma. \tag{2}$$

The integration is performed over the entire range of the parameters Γ. The traditional methods for computing z rely on Monte Carlo and Quasi Monte Carlo methods, where $y(t,\gamma)$ is sampled for a sequence of points in the parameter space.

Interpolatory parametric model order reduction techniques build a reduced model for (1) by interpolation in specific well chosen points in the parameter space. The advantage of such an approach is that (nonparametric) model reduction can be used in each of these points. The obtained reduced models are then joined in a parametric model or are glued together through interpolation [2] [4] [5] [3]. Other techniques are based on multivariate moment matching [11][9][10][7].

In this talk, the parameters are discretized by a Cartesian grid containing points $\gamma_1, \ldots, \gamma_N$. We build a parameter-free model for z obtained from Gaussian quadrature for (2). This leads to another single-input single-output system, but now with large block diagonal matrices whose diagonal blocks are $A(\gamma_j)$, $j = 1, \ldots, N$. The state vector of this system is represented by a tensor, stored in a low rank format, such as Tensor Trains or Hierarchical Tucker Tensors. We expect the explosion of unknowns due to the Cartesian grid to be compensated by the low rank structure of these tensors.

We then apply a two-sided moment matching method on this large system, using Padé via Lanczos [6] or the two-sided Arnoldi method. For moment matching at zero, we have to solve a large block diagonal linear system on each iteration. This is performed using a tensor Krylov method [8]. We thus have an inner-outer iteration scheme. We consider three scenarios for moment matching at zero: first, when the (inner) tensor Krylov iterations are preconditioned with $A(\gamma_0)^{-1}$ for a fixed parameter $\gamma_0 \in \Gamma$, then we show a connection with multivariate Padé approximation [9][10][7]. Interesting approximation properties can be proven for interpolation points selected from Gauss quadrature rules. In the second scenario, we perform many inner tensor Krylov iterations so that the obtained moments are interpolated accurately in the grid points in Γ. This then leads to an interpolatory reduced model. In the third scenario, we use an approximate inversion of $A(\gamma_0)$ instead of an exact inversion.

References

[1] S. Adhikari and M.I. Friswell. Random matrix eigenvalue problems in structural dynamics. *International Journal of Numerical Methods in Engineering*, 69:562–591, 2007.

[2] A.C. Antoulas, C.A. Beattie, and S. Gugercin. *Interpolatory model reduction of large-scale dynamical systems*, page 56. Springer-Verlag, 2010.

[3] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An empirical interpolation method: Application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339:667–672, 2004.

[4] U. Baur, P. Benner, C. Beattie, and S. Gugercin. Interpolatory projection methods for parameterized model reduction. *SIAM J. Sci. Comput.*, 33(5):2489–2518, 2011.

[5] T. Bui-Thanh, K. Willcox, and O. Ghattas. Model reduction for large scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computing*, 30(6):3270–3288, 2008.

[6] P. Feldman and R. W. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Trans. Computer-Aided Design*, CAD-14:639–649, 1995.

[7] L. Feng. Parameter independent model order reduction. *Mathematics and Computers in Simulation*, 68:221–234, 2005.

[8] D. Kressner and C. Tobler. Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1288–1316, 2011.

[9] Y.-T. Li, Z. Bai, and Y. Su. A two-directional Arnoldi process and its application to parametric model order reduction. *Journal of Computational and Applied Mathematics*, 226:10–21, 2009.

[10] Y.-T. Li, Z. Bai, Y. Su, and X. Zeng. Model order reduction of parameterized interconnect networks via a Two-Directional Arnoldi process. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 27(9):1571–1582, 2008.

[11] D. S. Weile, Michielssen E., E. Grimme, and K. Gallivan. A method for generating rational interpolant reduced order models of two-parameter linear systems. *Appl. Math. Let.*, 12:93–102, 1999.

Generic Rank-One Perturbations: Structure Defeats Sensitivity

Christian Mehl, Volker Mehrmann, Andre Ran and Leiba Rodman

Abstract

The behaviour of eigenvalues of matrices under perturbations is a frequently study topic in Numerical Linear Algebra. In particular, the study of low rank perturbations is well established and well understood by now. For example, in [1] it was shown that if $A \in \mathbb{C}^{n\times n}$ is a defective matrix having the eigenvalue λ with partial multiplicities $n_1 \geq \cdots \geq n_k$, i.e., having k Jordan blocks associated with λ with sizes $n_1, \ldots, n_k$, then applying a generic rank one perturbation results in a matrix having the eigenvalue λ with partial multiplicities $n_2 \geq \cdots \geq n_k$. As simple arguments show that the geometric multiplicity can only decrease by one if a rank-one perturbation is applied, we find that it is the largest Jordan block that is the most sensitive one to generic rank-one perturbations.

The picture changes drastically if structured matrices are considered and if the class of perturbation matrices is restricted to perturbations that preserve the original structure of the matrix. For example, consider *Hamiltonian matrices*, i.e., matrices $H \in \mathbb{C}^{2n\times 2n}$ satisfying $H^TJ + JH = 0$, where

$$J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}.$$

If we consider the Hamiltonian matrix $H = 0 \in \mathbb{C}^{2\times 2}$, then a generic rank-one perturbation would result in a 2×2 matrix having the eigenvalue 0 and a nonzero eigenvalue - an observation that is in accordance with the theory from [1] that (one of) the largest Jordan block(s) is destroyed, but the other Jordan block(s) will remains. On the other hand, a rank-one perturbation that is itself Hamiltonian necessarily has the form

$$\Delta H = uu^TJ = \begin{bmatrix} -u_1u_2 & u_1^2 \\ -u_2^2 & u_1u_2 \end{bmatrix}$$

for some vector $u = (u_1, u_2)^T \in \mathbb{C}^2$. It follows that $H + \Delta H = \Delta H$ is nilpotent for any choice of u and thus it has a Jordan block associated with $\lambda = 0$ of size two unless $u = 0$. Thus, incontrast to the unstructured case, a generic Hamiltonian rank-one perturbation results in a matrix, where the largest Jordan block has increased in size. From this point of view *structure defeats sensitivity* because the behaviour under perturbations of the largest (and most sensitive) Jordan block changes if structure-preservation is enforced.

The effects of structured rank-one perturbations become even more surprising when H-orthogonal matrices are considered. If $H \in \mathbb{C}^{n\times n}$ is a nonsingular symmetric matrix, then a matrix $U \in \mathbb{C}^{n\times n}$ is called *H-orthogonal* if $U^THU = H$. Considering the example

$$U = \begin{bmatrix} \lambda & 0 \\ 0 & -\lambda \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

for some $\lambda \in \mathbb{C} \setminus \{0\}$ it follows that $U^THU = H$, i.e., U is H-orthogonal. According to [1] a generic unstructured rank-one perturbation would result in a matrix having two arbitrary complex eigenvalues. On the other hand, a straightforward calculation shows that any H-orthogonal rank-one perturbation of U has the form

$$U\left(I_2 - 2\frac{2}{u^THu}uu^TH\right) = U\left(I_2 - \frac{1}{u_1u_2}\begin{bmatrix} u_1u_2 & u_1^2 \\ u_2^2 & u_1u_2 \end{bmatrix}\right) = \begin{bmatrix} 0 & \frac{u_1}{u_2}\lambda^{-1} \\ \frac{u_2}{u_1}\lambda & 0 \end{bmatrix}$$

which has the eigenvalues $+1$ and -1. Thus, all H-orthogonal rank-one perturbation of the original matrix have identical spectrum which is a rather surprising behaviour. Another surprising example is given by the H-orthogonal matrix

$$U = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

One can show that any H-orthogonal rank-one perturbation of U has the Jordan canonical form

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Again, all H-orthogonal rank-one perturbations of U have identical spectrum.

In the talk we will investigate the effects of structure-preserving rank-one perturbations of matrices that are structured with respect to an indefinite inner products with special emphasis on H-orthogonal and Hamiltonian matrices. We will present some general results on rank-one perturbations of structured matrices and by the help of those explain the surprising behaviour of structure-preserving perturbations of H-orthogonal and Hamiltonian matrices.

References

[1] J. Moro and F. Dopico. Low rank perturbation of Jordan structure. *SIAM J. Matrix Anal. Appl.*, 25:495–506, 2003.

Numerical Solution of Large Scale Parametric Eigenvalue Problems arising in the Analysis of Brake Squeal

Volker Mehrmann, Sarosh Quraishi and Christian Schröder

Abstract

We discuss model order reduction methods in the context of the simulation and accurate prediction of disc brake squeal. Brake squeal is a common problem in automobiles which occurs when the brakes are suddenly applied in a moving automobile and it is a significant problems in modern brake design. Physically, brake squeal can be interpreted as a self-excited vibration caused by fluctuations in the friction forces at the pad-rotor interface. This friction-induced self-excitation of a brake system is usually investigated based on the eigenvalues of linearized models. Even though the number of degrees of freedom and computational effort has constantly increased, there is still a lack of correspondence between experimental investigations and simulations. This work gives an overview of the finite element modeling (FEM) and analyzes the commonly used solution procedures for the high dimensional eigenvalue problem in the context of parametric model order reduction.

Solving high dimensional linear differential equations is a common problem within FEM. Usually in structural analysis, the unforced systems are considered in the form

$$M\ddot{q} + D\dot{q} + Kq = 0,$$

where M, D, $K \in \mathbb{R}^{n,n}$ are mass, damping and stiffness matrices, respectively, which are usually symmetric and positive semi-definite or definite. The vector valued function $q : \mathbb{R} \to \mathbb{R}^n$ contains the position variables associated with the degrees of freedom, arising from the finite element discretization.

In the analysis and simulation of brake squeal, however, the matrices D and K are typically non-symmetric and dependent on parameters such as the rotational speed of the brake disc, the braking pressure, the friction coefficient and other parameters. The non-symmetry arises from the fact that gyroscopic as well as circulatory forces have to be included as well as special parameter dependent damping models.

To be able to determine the stability behavior of a finite element (FE) model of the brake, right half plane eigenvalues of non-symmetric quadratic eigenvalue problems have to be calculated for many choices of parameters. We will discuss a new parameter dependent subspace projection approach, where the subspaces are determined from non-symmetric problems evaluated at many different parameter values and we will demonstrate that this approach is more accurate than traditionally used algorithms in industry.

Computing the Exponential of a Large Block Triangular Block Toeplitz Matrix

D.A. Bini, S. Dendievel, G. Latouche and B. Meini

Abstract

The Erlangian approximation of Markovian fluid queues leads to the problem of computing the exponential of an upper block triangular, block Toeplitz matrix [2]

$$U = \begin{bmatrix} U_0 & U_1 & \dots & U_{\ell-1} \\ & U_0 & \ddots & \vdots \\ & & \ddots & U_1 \\ 0 & & & U_0 \end{bmatrix},$$

where U_i, $i = 0, \dots, \ell-1$, are $m \times m$ matrices. The block size ℓ of U may be huge, since a larger ℓ leads to a better Erlangian approximation, while the size m of the blocks is generally small. The matrix U is a subgenerator, i.e., it has negative diagonal entries, nonnegative off-diagonal entries, and the sum of the entries on each row in nonpositive. The matrix exponential $A = \mathrm{e}^U$ is still an upper block triangular, block Toeplitz matrix; in particular, the diagonal blocks of A are the matrices e^{U_0}. The matrix A is nonnegative and substochastic.

We propose some numerical methods for computing the exponential of the matrix U, that exploit the block triangular block Toeplitz structure. Unlike the general methods, our algorithms allow to deal with very large sizes.

Two numerical methods rely on the property that a block z-circulant matrix can be block diagonalized by means of Fast Fourier Transforms [1]. Therefore the computation of the exponential of an $n \times n$ z-block circulant matrix with $m \times m$ blocks can be reduced to the computation of n exponentials of $m \times m$ matrices.

The idea of the first method is to approximate $A = \mathrm{e}^U$ by the exponential of the block ϵ-circulant matrix

$$U_\epsilon = \begin{bmatrix} U_0 & U_1 & \dots & U_{\ell-1} \\ \epsilon U_{\ell-1} & U_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & U_1 \\ \epsilon U_1 & \dots & \epsilon U_{\ell-1} & U_0 \end{bmatrix},$$

where $\epsilon \in \mathbb{C}$ and $|\epsilon|$ is sufficiently small. An error analysis is performed, that allows to choose the value of ϵ. The computation of e^{U_ϵ} is reduced to the computation of ℓ exponentials of $m \times m$ matrices.

In the second approach the matrix U is embedded into a $K \times K$ block circulant matrix C_K, where K is sufficiently large, and an approximation of e^U is obtained from a suitable submatrix of e^{C_K}. The computation of e^{C_K} is reduced to the computation of K exponentials of $m \times m$ matrices. In practical computations, we start with $K = 2\ell$, and the value of K is doubled until a certain condition is satisfied.

Another numerical method consists in specializing the shifting and Taylor series method of [3]. The block triangular Toeplitz structure is exploited in the FFT-based matrix multiplications involved in the algorithm, leading to a reduction of the computational cost.

Theoretical and numerical comparisons among the three numerical methods are presented.

References

[1] D. Bini. Parallel solution of certain Toeplitz linear systems. *SIAM J. Comput.*, 13 (1984), no. 2, 268–276.

[2] N.J. Higham. *Functions of matrices. Theory and computation*. SIAM, Philadelphia, PA, 2008.

[3] J. Xue, Q. Ye. Computing exponentials of essentially non-negative matrices entrywise to high relative accuracy. *Math. Comp.*, 82 (2013), no. 283, 1577–1596.

Numerical Optimization of Eigenvalues of Hermitian Matrix-Valued Functions

Emre Mengi, Emre Alper Yildirim and Mustafa Kilic

Abstract

Optimizing a prescribed eigenvalue of a differential operator or a matrix-valued function depending on parameters has a long history dating back to Euler. The prescribed eigenvalue of interest typically exhibits not only Lipschitz continuity, but also remarkable analyticity properties. We present a numerical approach [3], based on these analytical properties, for unconstrained optimization of a prescribed eigenvalue of a Hermitian matrix-valued function depending on a few parameters analytically. In the second part, we present a numerical approach [4] for optimization problems with linear objectives and an eigenvalue constraint on the matrix-valued function. The common theme for numerical approaches is the use of global under-estimators, called support functions, for eigenvalue functions, that are built on the variational properties of eigenvalues over the space of Hermitian matrices.

The unconstrained eigenvalue optimization problem is

$$\min_{\omega\in\mathbb{R}^d,\omega\in\mathcal{B}_d} \lambda_j(\mathcal{A}(\omega))$$

where $\mathcal{A}(\omega):\mathbb{R}^d\to\mathbb{C}^{n\times n}$ is a Hermitian and an analytic function of ω, λ_j denotes the jth largest eigenvalue, and $\mathcal{B}_d$ is a given box in $\mathbb{R}^d$. Due to the classical results by Rellich [1], the eigenvalues of $\mathcal{A}(\omega)$ has a particular ordering $\tilde{\lambda}_1(\omega),\ldots,\tilde{\lambda}_n(\omega)$ so that each eigenvalue $\tilde{\lambda}_\ell(\omega)$ is analytic along every line in $\mathbb{R}^d$. Applications of Taylor's theorem combined with Rellich's results yield the support function

$$s(\omega;\omega_k) := \lambda_j(\mathcal{A}(\omega_k)) + \left(\min_{\ell=1,\ldots,n}\nabla\tilde{\lambda}_\ell(\omega_k)^T(\omega-\omega_k)\right) + \frac{\gamma}{2}\|\omega-\omega_k\|^2.$$

about a given ω_k satisfying $s(\omega;\omega_k)\le\lambda_j(\mathcal{A}(\omega))$ for all ω, and $s(\omega_k;\omega_k)=\lambda_j(\mathcal{A}(\omega_k))$. Here γ is any real scalar such that $\lambda_{\min}\left(\nabla^2\tilde{\lambda}_\ell(\omega)\right)\ge\gamma,\quad\forall\omega\in\mathcal{B}_d,\quad\ell=1,\ldots,n$. For the maximal eigenvalue function $\lambda_1(\mathcal{A}(\omega))$ this support function takes the more plausible quadratic form

$$s(\omega;\omega_k) := \lambda(\omega_k) + \nabla\lambda_1(\mathcal{A}(\omega_k))^T(\omega-\omega_k) + \frac{\gamma}{2}\|\omega-\omega_k\|^2$$

about any ω_k such that $\lambda_1(\mathcal{A}(\omega_k))$ is simple. In this case γ is any lower bound for $\lambda_{\min}\left(\nabla^2\lambda_1(\mathcal{A}(\omega))\right)$ for all $\omega\in\mathcal{B}_d$ such that $\lambda_1(\mathcal{A}(\omega))$ is simple. The numerical scheme is based on the approximation of the eigenvalue function with the model $s(\omega) := \max_{k=0,\ldots,p} s(\omega;\omega_k)$ in terms of the support functions about $\omega_0,\ldots,\omega_p$. The point ω_{p+1} is defined to be any global minimizer of $s(\omega)$, and the model $s(\omega)$ is refined with the incorporation of $s(\omega;\omega_{p+1})$. This leads us to a globally convergent algorithm, i.e., every convergent subsequence of $\{\omega_k\}$ converges to a global minimizer of $\lambda_j(\mathcal{A}(\omega))$. The rate of convergence appears linear in practice. The practicality of the algorithm relies on two issues: **(1)** *numerical computation of a global minimizer of* $s(\omega)$ - support function ideas have already been pursued by the global optimization community, here we benefit from existing schemes especially the one due to Breiman and Cutler [2]; **(2)** *analytical or numerical deduction of* γ - we show that analytical deduction of γ is possible in various occasions for the maximal eigenvalue function (more generally the sum of the j largest eigenvalues), in some other cases it is possible to deduce γ numerically.

The constrained eigenvalue optimization problem is

$$\max_{\omega \in \mathbb{R}^d} \; c^T \omega \quad \text{subject to} \quad \lambda_n(\mathcal{A}(\omega)) \leq 0$$

where $\mathcal{A}(\omega)$ is as in the unconstrained eigenvalue optimization problem, $c \in \mathbb{R}^d$ is given, and we assume the feasible set is bounded. For instance all of the pseudospectral functions, e.g., those concerning the standard, polynomial eigenvalue problems and those occurring from the delay systems, fit into this framework. In this setting, given any γ that is an upper bound for $\lambda_{\max}\left(\nabla^2 \lambda_n(\mathcal{A}(\omega))\right)$ for all ω, the use of the (upper) support function

$$s(\omega; \omega_k) := \lambda(\omega_k) + \nabla \lambda_n(\mathcal{A}(\omega_k))^T (\omega - \omega_k) + \frac{\gamma}{2} \|\omega - \omega_k\|^2$$

about ω_k such that $\lambda_n(\mathcal{A}(\omega_k))$ is simple, satisfying $s(\omega; \omega_k) \geq \lambda_n(\mathcal{A}(\omega))$ for all ω and $s(\omega_k; \omega_k) = \lambda_n(\mathcal{A}(\omega_k))$, yields the convex program

$$\max_{\omega \in \mathbb{R}^d} \; c^T \omega \quad \text{subject to} \quad s(\omega; \omega_k) \leq 0.$$

The feasible set for this convex program is a subset of that of the constrained eigenvalue optimization problem. Consequently, starting from a feasible point ω_0 for the original problem, we generate a feasible sequence $\{\omega_k\}$ where ω_{k+1} is the unique local maximizer of the convex program above, which can be determined analytically. The algorithm can be viewed as a fixed-point iteration. Unlike the unconstrained counterpart, this algorithm converges locally, but can be neatly analyzed. For instance, we prove its linear rate of convergence; indeed the main factor affecting the speed of convergence is the eigenvalue distribution of a projected Hessian at the converged local maximizer, in particular the closer are the eigenvalues of the projected Hessian to γ, the faster is the convergence.

References

[1] F. Rellich. *Perturbation Theory of Eigenvalue Problems*. Gordon and Breach, 1969.

[2] L. Breiman and A. Cutler. A deterministic algorithm for global optimization. *Math. Program.*, 58(2):179:199, February 1993.

[3] E. Mengi, E.A. Yildirim and M. Kilic. Numerical optimization of eigenvalues of Hermitian matrix functions. *SIAM J. Matrix Anal. Appl.*, submitted.

[4] E. Mengi. A support based algorithm for optimization with eigenvalue constraints. *math arXiv:13101563*.

On the convergence of QOR and QMR Krylov methods for solving linear systems

Jurjen Duintjer Tebbens and Gerard Meurant

Abstract

We consider the problem of solving linear systems $Ax = b$ where A is a square nonsingular matrix of order n with real or complex entries and b is a vector of length n. The most popular iterative methods for solving such a possibly nonsymmetric linear system are Krylov methods. In [6] it is shown that most Krylov methods can be described as quasi-orthogonal (QOR) or quasi-minimum residual (QMR) methods.

Many numerical methods that can be classified as QOR or QMR methods have been proposed over the years. Generally they come into QOR/QMR pairs. Probably the most famous ones are the FOM (Saad) and GMRES (Saad and Schultz) algorithms (which are in fact true OR and MR methods since the matrix V representing the basis of the Krylov subspace is computed to be orthonormal). The matrices V and the upper Hessenberg matrix $H = V^*AV$ are computed using the Arnoldi process.

Another famous pair is BiCG/QMR. BiCG and QMR (Freund and Nachtigal) use a basis represented by V that is bi-orthogonal to a basis W for the Krylov subspace constructed on (A^*, b). QMR minimizes the quasi-residual norm. Note that in this presentation we will use the acronym QMR in two different ways. The first one is to denote the general class of quasi-minimum residual methods and the second one is related to the particular method introduced by Freund and Nachtigal.

A third pair of methods is Hessenberg/CMRH. CMRH, introduced by Sadok, is a QMR method which uses the Hessenberg basis computed with an LU factorization with partial pivoting of the Krylov matrix; see [10]. The Hessenberg method is the corresponding QOR method.

Other examples include truncated methods and some restarted methods as long as they build and use a basis of the Krylov subspace.

In this talk we are interested in extending the results obtained in [7, 1, 8], [2, 3], [5], [4], [9] for FOM/GMRES to other Krylov methods with non-orthonormal bases. Rather than doing this method by method we will adopt the general framework of [6] and we will not consider particular implementations. In [7, 1, 8] and later on in [2, 3] parametrizations of the classes of matrices (with a prescribed spectrum) and right-hand sides giving prescribed GMRES residual norms were given as well as practical ways of constructing such matrices and right-hand sides; see also [9]. In [5] and [4] closed-form expressions of the GMRES residual norms as functions of the eigenvalues and eigenvectors as well as the right-hand side were provided. Our aim is to show that most of these results known for FOM/GMRES can be extended to general QOR/QMR methods.

The contents of the presentation are the following:

- We first describe the general framework for QOR and QMR Krylov methods using basis vectors which are the columns of a non-orthonormal nonsingular matrix V. We show expressions for principal submatrices of the upper Hessenberg matrix H as functions of entries of the change of basis upper triangular matrix U defined by $K = VU$, K being the Krylov matrix.
- We show that the QOR residual norms can be read from the first row of U^{-1} and we give an expression for the QMR quasi-residual norms as a function of the entries of the matrix U^*U.

- Two parametrizations (or factorizations) of the matrix A and of the right-hand side b are provided corresponding to those in [1] and [2].

- We consider the problem of constructing a matrix (with a prescribed spectrum) and a right-hand side yielding prescribed QOR residual norm or QMR quasi-residual norm convergence curves for a given QOR or QMR method. Depending on the method this amounts to construct a matrix H with a particular non-zero structure. We restrict ourselves to a banded structure in the upper part of H.

- We give expressions for the QMR quasi-residual norms as a function of the eigenvalues and eigenvectors of A, the right-hand side b and the matrix V of the basis vectors. It is a generalization of the result in [4].

- We study some relationships between general QOR/QMR methods and FOM/GMRES.

- Finally we describe two numerical experiments in which we construct matrices with a given spectrum and with a prescribed residual norm convergence curve for BiCG and for the Hessenberg method.

Hence the situation for these QOR/QMR methods is the same as for FOM/GMRES. One can obtain the same convergence curves with matrices having arbitrary prescribed spectra and the (quasi-) residual norms depend in an intricate way on the eigenvalues, the eigenvectors, the right-hand side and the basis vectors.

References

[1] M. Arioli, V. Pták and Z. Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT Numerical Mathematics, v 38 n 4 (1998), pp. 636–643.

[2] J. Duintjer Tebbens and G. Meurant, *Any Ritz value behavior is possible for Arnoldi and for GMRES*, SIAM J. Matrix Anal. Appl., v 33 n 3 (2012), pp. 958–978.

[3] J. Duintjer Tebbens and G. Meurant, *Prescribing the behaviour of early terminating GMRES and Arnoldi iterations*, to appear in Numerical Algorithms, (2013).

[4] J. Duintjer Tebbens and G. Meurant, *GMRES convergence and the Jordan canonical form*, in preparation, (2013).

[5] J. Duintjer Tebbens, G. Meurant, H. Sadok and Z. Strakoš, *On investigating GMRES convergence using unitary matrices*, submitted, (2013).

[6] M. Eiermann and O.G. Ernst, *Geometric aspects in the theory of Krylov subspace methods*, Acta Numerica, v 10 n 10 (2001), pp. 251–312.

[7] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent advances in iterative methods, G.H. Golub, A. Greenbaum and M. Luskin, eds., Springer, (1994), pp. 95–118.

[8] A. Greenbaum, V. Pták and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., v 17 (1996), pp. 465–469.

[9] G. Meurant, *GMRES and the Arioli, Pták and Strakoš factorization*, BIT Numerical Mathematics, v 52 n 3 (2012), pp. 687–702.

[10] H. Sadok, *CMRH: A new method for solving nonsymmetric linear system based on the Hessenberg reduction algorithm*, Numer. Algo., v 20 (1999), pp. 303–321.

A Posteriori Error Estimates for hp-Adaptive Approximations of Non-selfadjoint PDE Eigenvalue Problems

Stefano Giani, Luka Grubišić, Agnieszka Międlar and Jeffrey S. Ovall

Abstract

Let us consider the following problem in bounded, polygonal domain $\Omega \subset \mathbb{R}^2$:

Find $(\lambda, \psi) \in \mathbb{C} \times H_0^1(\Omega)$ *such that*

$$\mathcal{A}\psi := -\nabla \cdot A\nabla\psi + b \cdot \nabla\psi + c\psi = \lambda\psi \ , \tag{1}$$

with real-valued $A \in \left[L^\infty(\Omega)\right]^{2\times 2}$, $b \in \left[L^\infty(\Omega)\right]^2$ *with* $\nabla \cdot b \in L^\infty(\Omega)$, *and* $c \in L^\infty(\Omega)$.

We introduce new residual a posteriori eigenvalue/eigenvector error estimates based on Kato's square root theorem [6] for spectral approximations of diagonalizable non-selfadjoint differential operators of convection-diffusion-reaction type with real spectrum. Problem (1) provides an example of a more general class of non-selfadjoint eigenvalue problems in Hilbert space for which a Riesz basis can be constructed from associated eigenvectors [3, 5]. Under, the so called *Condition* $\mathfrak{S}$, i.e.,

Definition 1 (*Condition* $\mathfrak{S}$)**.** *The operator* $\mathcal{A}$ *satisfies the condition* $\mathfrak{S}$, *where* $\mathfrak{S}$ *stands for the square root, if it is of the form (1) and is diagonalizable, i.e.,* $\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}$ *with the normal operator* $\mathcal{H}$ *being selfadjoint and positive definite.*

For operators which satisfy condition $\mathfrak{S}$, we define the square root operator

$$\mathcal{A}^{1/2} = \mathcal{X}\mathcal{H}^{1/2}\mathcal{X}^{-1},$$

which domain is the same as the domain of the abstract bilinear form $B(\cdot,\cdot)$ defining the operator $\mathcal{A}$ [1, 2]. This guarantees the existence of constants c_K, c_K^*, C_K and C_K^* such that

$$\begin{aligned} c_K\|\phi\|_1 &\le \|\mathcal{A}^{1/2}\phi\| \le C_K\|\phi\|_1, \qquad \phi \in H_0^1(\Omega), \\ c_K^*\|\phi\|_1 &\le \|\mathcal{A}^{*1/2}\phi\| \le C_K^*\|\phi\|_1, \qquad \phi \in H_0^1(\Omega) \ . \end{aligned}$$

Exploiting aforementioned results and residual eigenvector/eigenvalue error estimates, we obtain the following *Bauer–Fike* type estimate:

Theorem 1 (Bauer–Fike Estimate)**.** *Let* $\widehat{\mu} > 0$ *and* $\widehat{\psi} \in H_0^1(\Omega)$, $\|\widehat{\psi}\| = 1$ *be given and let* $\mathcal{A}$ *satisfy the condition* $\mathfrak{S}$. *Then*

$$\min_{\xi\in\mathrm{Spec}(\mathcal{A})} \frac{|\widehat{\mu}-\xi|}{\sqrt{\widehat{\mu}\xi}} \le \frac{\kappa(\mathcal{X})}{c_K^*}\frac{\|\mathfrak{r}(\widehat{\mu})[\widehat{\psi},\cdot]\|_{-1}}{\sqrt{\widehat{\mu}}}.$$

Moreover, we present an hp-adaptive finite element algorithm with residual a posteriori hp-finite element error estimates which are of higher order in the residual norm $\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi},\cdot]\|_{-1}$.

Theorem 2. *Let* $(\widehat{\lambda}, \widehat{\psi}, \widehat{\psi^\star}) \in \mathbb{R} \times V_h^p \times V_h^p$ *be an eigentriple of discretized eigenproblem (1). There is a constant* C *depending only on shape-regularity parameter* γ *for which*

$$\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi},\cdot]\|_{-1} \le C\widehat{\lambda}\eta(\widehat{\psi}) \quad \text{and} \quad \|\mathfrak{r}(\widehat{\lambda})[\cdot,\widehat{\psi^\star}]\|_{-1} \le C\widehat{\lambda}\eta^\star(\widehat{\psi^\star}) \ .$$

This result together with Theorem 1 allow to formulate the following key a posteriori error estimation results:

Theorem 3. *Let the operator $\mathcal{A}$ defined by the variational form $B(\cdot,\cdot)$ satisfy condition $\mathfrak{S}$ and let $(\widehat{\lambda},\widehat{\psi},\widehat{\psi^\star})\in\mathbb{R}\times V_h^p\times V_h^p$ be an eigentriple of discretized eigenproblem such that $(\widehat{\psi},\widehat{\psi^\star})\neq 0$. Then there exists an eigenvalue λ of $\mathcal{A}$ such that*

$$\frac{|\widehat{\lambda}-\lambda|}{\widehat{\lambda}}\leq C\widehat{\lambda}\eta(\widehat{\psi})\eta^\star(\widehat{\psi^\star})+o(\widehat{\lambda}\eta(\widehat{\psi})\eta^\star(\widehat{\psi^\star}))\ .$$

Furthermore, the flowing estimates hold

$$\|\psi-\widehat{\psi}\|_1\leq C\widehat{\lambda}\,\eta(\widehat{\psi}),\qquad \|\psi^\star-\widehat{\psi^\star}\|_1\leq C\widehat{\lambda}\,\eta^\star(\widehat{\psi^\star})\ .$$

We notice that Theorem 3 is the second order eigenvalue estimate which requires some additional assumptions on the convergence of Galerkin eigenvalue approximations, however, we can guarantee the first order convergence of the eigenvalues, namely:

For each ν, there exists an eigenvalue λ such that

$$|\widehat{\lambda}_\nu-\lambda|\leq\frac{\kappa(\mathcal{X})}{c_K^*}\ C\ \sqrt{\lambda}\ \widehat{\lambda}_\nu\eta(\widehat{\psi}_\nu)\ .$$

For detailed analysis we refer the reader to [4].

References

[1] P. Auscher, S. Hofmann, M. Lacey, A. McIntosh, and P. Tchamitchian, *The solution of the Kato square root problem for second order elliptic operators on $\mathbb{R}^n$*, Ann. of Math. (2) **156** (2002), no. 2, 633–654.

[2] P. Auscher and P. Tchamitchian, *Square roots of elliptic second order divergence operators on strongly lipschitz domains: L^2 theory*, J. Anal. Math. **90** (2003), no. 1, 1–12.

[3] E. B. Davies, *Linear operators and their spectra*, Cambridge Studies in Advanced Mathematics, vol. 106, Cambridge University Press, Cambridge, 2007.

[4] S. Giani, L. Grubišić, A. Międlar, and J. S. Ovall, *Robust estimates for hp-adaptive approximations of non-self-adjoint eigenvalue problems*, Preprint 1008, DFG Research Center Matheon, 2013.

[5] I. C. Gohberg and M. G. Kreĭn, *Introduction to the theory of linear nonselfadjoint operators*, Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs, Vol. 18, American Mathematical Society, Providence, R.I., 1969.

[6] T. Kato, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, Reprint of the 1980 edition.

Resurrecting the Symmetric Generalized Matrix Eigenvalue Problem

Cleve Moler

Abstract

I want to try to stimulate renewed investigation of the eigenvalue problem

$$Ax = \lambda Bx$$

for symmetric dense matrices A and B of modest order. Here B is positive definite, but nearly singular. I believe that we do not yet have completely satisfactory algorithms for this problem. LAPACK and MATLAB offer the choice between the inversion of the ill-conditioned Cholesky factors of B and the QZ algorithm that destroys symmetry. I recently found an unfinished paper that I was working on with Jim Wilkinson in 1986 that proposes a new algorithm involving double diagonalization using square root free Givens-like transformations. The resulting program may not be competitive with modern codes in speed, but does offer advantages in accuracy.

A Multigrid Arnoldi Method for Eigenvalues

Ron Morgan

Abstract

Introduction

We consider Krylov subspace methods for computing eigenvalues and eigenvectors of large, possibly nonsymmetric matrices. Sorensen's Implicitly Restarted Arnoldi Method [3] (IRAM) was a leap forward for finding several eigenvalues. However, many eigenvalue problems are still challenging. For problems coming from discretization of differential equations, fine discretizations lead to large matrices, wide spectra and difficult computations. It is possible that the next leap will involve multigrid. An approach is proposed here. First we give an example illustrating slow convergence.

Example 1. We consider a matrix from finite difference discretization of the 2-D convection-diffusion equation $-u_{xx} - u_{yy} + 10u_x = \lambda u$ on the unit square. We let $h = \frac{1}{702}$, leading to a matrix of dimension $n = 491,401$. Eigenvalues range from $9.0 * 10^{-5}$ to 8 with difficult near-multiplicity. A restarted Arnoldi method equivalent to IRAM is run with the goal of finding the eight smallest eigenpairs to residual norms below 10^{-8}. We use subspaces of maximum size 30 and restart with 15 Ritz vectors (denoted by Arn(30,15)). This method takes 1594 cycles or 23,925 mat-vecs.

To see a version of this abstract with graphs:
https://bearspace.baylor.edu/Ronald_Morgan/www/MorganHH14.pdf

Multigrid Arnoldi

We wish to improve convergence by taking advantage of coarse grids. Multigrid has been proposed for eigenvalue problems in a number of different ways, but there does not seem to be an established method. Some eigenvalue multigrid methods only compute one eigenvalue at a time. Also, generally the methods proposed are subject to the same limitations as multigrid for linear equations: multigrid does not converge for convection-diffusion with too much convection and for many indefinite problems.

We propose a Krylov multigrid method that can compute many eigenvalues simultaneously and can solve problems for which standard multigrid fails. Most significantly, it has potential to dramatically improve computation of eigenvalues and eigenvectors. We present the method here on two grids.

Two-grid Arnoldi

1. *Run restarted Arnoldi on coarse grid.*
2. *Use coarse grid eigenvectors to create approximate eigenvectors on fine grid (splines).*
3. *Improve approximate eigenvectors on fine grid with the Arnoldi-E [1] method.*

The Arnoldi-E method mentioned in Step 3 builds a Krylov subspace and then appends approximate eigenvectors at the end of the cycle. This allows for approximate eigenvectors to be input at the beginning of the method.

Example 1 (cont.). We apply the Two-grid Arnoldi approach to the eigenvalue problem in Example 1. For the coarse grid, we use discretization size of $\frac{1}{351}$, so the number of grid points and dimension

of the matrix is $350^2 = 122,500$. This is about one-fourth of the dimension of the fine grid matrix. Only 447 cycles of Arn(30,15) are required to find the smallest eight eigenvalues to accuracy of residual norm below 10^{-8}. This is under a third of the cycles for the larger problem. However, the cost is actually much less than this, because with a smaller matrix and shorter vectors, the cost per cycle is about one-fourth as much. In this experiment, we run 500 cycles on the coarse grid, then move the coarse grid eigenvectors to the fine grid and improve them there. The Arnoldi-E method needs only 34 cycles on the fine grid for the eigenvectors to reach the desired level. To better compare this two-grid approach, we multiply the number of coarse grid cycles by one-fourth and add this to the number of fine grid cycles. This gives 159 fine-grid-equivalent cycles compared to the 1594 for regular Arnoldi.

We note that results for the new Two-grid Arnoldi may not be as impressive relative to regular Arnoldi when the coarse grid does not give accurate approximations for the fine grid. This is particulary likely if even the fine grid is fairly coarse. Also note that it is possible to use more than two grids.

Fine Grid Convergence Theory

Arnoldi-E does not always work well at improving approximate eigenvectors. There is something special about the approximate eigenvectors that come from the coarse grid. We give an example with symmetric matrix of size $n = 2047$ from a 1-D diffusion equation. We compare Arnoldi-E for improving approximate eigenvectors that come from a coarse grid with $n = 511$ versus from random perturbation of the true fine grid eigenvectors. The eventual convergence rate is roughly four times faster for improving the coarse grid vectors. We will discuss this convergence in the context of Krylov properties of restarted Arnoldi [1, 2], however here we will have near-Krylov subspaces [4].

Deflating Linear Equations

If time permits, we will discuss a two-grid approach for solving linear equations. Approximate eigenvectors from the coarse grid can be used to deflate eigenvalues for the linear equations. This can potentially apply the power of multigrid to problems for which regular multigrid methods fail. There is also potential for combining eigenvalue deflation with multigrid preconditioning.

References

[1] R. B. Morgan. On restarting the Arnoldi method for large nonsymmetric eigenvalue problems. *Math. Comp.*, 65:1213–1230, 1996.

[2] R. B. Morgan. GMRES with deflated restarting. *SIAM J. Sci. Comput.*, 24:20–37, 2002.

[3] D. C. Sorensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM J. Matrix Anal. Appl.*, 13:357–385, 1992.

[4] G. W. Stewart. Backward error bounds for approximate krylov subspaces. *Linear Algebra Appl.*, 340:81–86, 2002.

Multilevel Krylov Methods

Yogi A. Erlangga and Reinhard Nabben

Abstract

Preconditioned Krylov subspace methods are among the most efficient tools for solving large linear systems. But beside the traditional preconditioner there are also other techniques to accelerate the speed of convergence of a Krylov subspace method, namely deflation or augmentation techniques. Here we explain these techniques in detail with the help of projections. Surprisingly, similar projection techniques are known in the fields of domain decomposition and multigrid methods.

Based on our detailed analysis of deflation methods we develop a new multilevel deflation method or a multilevel Krylov method (MK-method). The basic idea of this type of method is to precondition a flexible Krylov subspace method with a preconditioner that shift some eigenvalues to an a priori fixed constant. The shifting of the eigenvalues is similar to projection type or deflation type methods and uses the solution of subspace or coarse level systems.

In contrast to multigrid methods or Krylov methods preconditioned by multigrid, the subspace systems are solved approximatively by a few steps of the flexible Krylov method again. In contrast to most deflation and augmentation methods no approximate eigenvectors are used.

The concept of our MK-methods yields to a new family of multilevel methods that consists of several ingredients that can be chosen independently: a traditional preconditioner; a flexible Krylov method; the multilevel structure, subspace (coarse gird) systems; restrictions, prolongations, deflation vectors.

Several theoretical and numerical results show the high potential of our multilevel Krylov methods.

We have done several numerical examples with the flexible GMRES method. It turns out that these MK methods work very well for convection diffusion equations and the Helmholtz equation. The convergence can be made almost independent of grid size h and also only mildly dependent of the wavenumber k for the Helmholtz equation.

Numerical results for an MK method using the flexible CG method (MK-CG method) for symmetric positive definite systems are given also. For the MK-CG method we establish a convergence bound.

Recently, we were able to prove some more theoretical results. With the help of Fourier analysis we are able to derive formulas for the eigenvalues of the MK operator applied to the 1D Helmholtz equation (joint work with Luis Garcia Ramos). These formulas show the extremely nice clustering of the eigenvalues. For MK-methods using the flexible GMRES method we proved convergence as long as the subsystems are solved up to a certain tolerance (joint work with R. Kehl and D.B. Szyld).

We also combine the multilevel Krylov idea with the algebraic way of choosing the subspace systems and the restriction and prolongation operators. The resulting method is called algebraic multilevel Krylov method or AMK method. We present different AMK methods that differ in the choice of the algebraic components. First, we use the classical Ruge and Stüben approach. Then, we present a agglomeration-based technique. Both methods are tested for various matrices arising from discretization of 2D diffusion and convection-diffusion equation as well as for several matrices taken from the matrix-market collections. The numerical results show that the AMK methods work as well as the geometric MK methods. For the convection-diffusion equations the AMK methods lead to better convergence rates than the original AMG method by Ruge and Stüben.

A Numerical Linear Algebraic Approach to Compact Multi-Frame Blind Deconvolution

James G. Nagy, Stuart Jefferies and Helen Schomburg

Abstract

We consider a multi-frame blind deconvolution (MFBD) problem that requires solving a large-scale nonlinear least squares problem of the form

$$\min_{\boldsymbol{x},\boldsymbol{y}_j} \sum_{j=1}^{J} \|\boldsymbol{A}(\boldsymbol{y}_j)\boldsymbol{x} - \boldsymbol{b}_j\|_2^2 \,, \tag{1}$$

where $\boldsymbol{b}_j$, $j = 1, 2, \ldots, N$, are observed images of an unknown object $\boldsymbol{x}$. The unknown $\boldsymbol{y}_j$ defines the linear image formation process, which depends on the conditions in which the data frame $\boldsymbol{b}_j$ was obtained. The term *blind* refers to the fact that the parameters $\boldsymbol{y}_j$ defining the matrices $\boldsymbol{A}_j = \boldsymbol{A}(\boldsymbol{y}_j)$ are unknown.

Standard algorithms to solve this problem usually rely on obtaining many frames of data, but they do not exploit the fact that there is likely to be substantial redundant information in the measured images. Recently, Hope and Jefferies [1] argued that ratios of the Fourier spectra of various images can be used to model the inherent correlations in the data, and they proposed to reduce the MFBD problem (1) to one that involves only a small number ($K \ll J$) of *control frames*. The reduced problem is referred to as compact multi-frame blind deconvolution (CMFBD).

We recast this application into a numerical linear algebra framework, which helps in the development of efficient solvers, and it opens the door to a wider class of problems on which the technique can be applied. Specifically, the application requires solving two problems. The first is a structured subset selection problem where we need to find a small subset of images from $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_J$ that contains all relevant information from the full dataset. The second problem involves reducing the MFBD problem (1) to a weighted nonlinear least squares problem

$$\min_{\boldsymbol{x},\boldsymbol{y}_k} \sum_{k=1}^{K} \left\| \boldsymbol{W}\boldsymbol{A}(\boldsymbol{y}_k)\boldsymbol{x} - \tilde{\boldsymbol{b}}_k \right\|_2^2 . \tag{2}$$

We show that this can be done very efficiently by exploiting structure of the matrices $\boldsymbol{A}(\boldsymbol{y}_j)$, and through the judicious use of Givens rotations.

References

[1] D. Hope and S. Jefferies, Compact multiframe blind deconvolution, Opt. Lett. 36, 867-869 (2011).

Using Zolotarev's high-order Rational Approximations for Computing the Polar, Symmetric Eigenvalue and Singular Value Decompositions

Yuji Nakatsukasa and Roland W. Freund

Abstract

The spectral divide-and-conquer algorithms for computing the symmetric eigenvalue decomposition and the SVD developed in [4] are near-optimal in both arithmetic and communication costs, and are crucially built on computing the polar decomposition $A = U_pH$. For the polar decomposition the QDWH iteration [3] is used, which requires at most six iterations to converge for any practical matrix in double precision arithmetic and uses the building blocks of just matrix multiplication and QR and Cholesky factorizations.

Following this framework we propose an algorithm for the polar decomposition, which is a higher-order variant of QDWH. The underlying theory is rational approximation to the scalar sign function on a union of intervals $[-1,-\epsilon]\cup[\epsilon,1]$, since denoting by $A = U\Sigma V^*$ the SVD of a full column rank A, the unitary polar factor is $U_p = UV^* = U\mathrm{sign}(\Sigma)V^*$. The key idea is to use the best rational approximation due to Zolotarev in 1877 [1]. We form the type $(m, m-1)$ best rational approximant $r_1(x) \in \mathcal{R}_{m,m-1}$ on $[-1,-\epsilon]\cup[\epsilon,1]$, and then form another best rational approximant $r_2(x) \in \mathcal{R}_{m,m-1}$ on a smaller interval $[-1,-\tilde{\epsilon}]\cup[\tilde{\epsilon},1]$ where $\tilde{\epsilon}(>\epsilon)$ is determined from ϵ and m. We run k iterations of this process to obtain $r_i(x) \in \mathcal{R}_{m,m-1}$ for $i = 1,\dots,k$. Then the composed function $r(x) = r_k(r_{k-1}(\cdots(r_1(x))$ is of type (m^k, m^k-1), an exponentially growing degree. Crucially, $r(x)$ happens to be Zolotarev's best rational approximant on the initial interval $[-1,-\epsilon]\cup[\epsilon,1]$ of this type. This optimality can be verified by counting the number of equioscillation points, a unique characterization of best rational approximants.

We are thereby able to form a very high degree (m^k) best rational approximant via composing k low degree (m) rational functions, which is attractive both for speed, because evaluating the latter needs $O(mkn^3)$ flops whereas the former requires $O(m^kn^3)$ flops, and for stability, because low-degree Zolotarev functions can be computed accurately at a matrix argument.

Some of the previously proposed algorithms are special cases of this framework: the QDWH algorithm uses $m = 3$, and the scaled Newton iteration with the scaling by Byers and Xu [2] uses $m = 2$. Previous experiments have suggested that the number of iterations needed by these algorithms for numerical convergence is bounded: six iterations for QDWH, which is a type $(3^6, 3^6-1) = (729, 728)$ Zolotarev best approximant, and nine iterations for scaled Newton, a type $(2^9, 2^9-1) = (512, 511)$ Zolotarev. This observation can be explained by the fact that a Zolotarev approximant of type $(d, d-1)$ with $d \geq 350$ has approximation error smaller than 10^{-15} for any $\epsilon > 10^{-15}$.

The algorithm we propose chooses $m = 19$ or smaller, where 19 is the smallest integer whose square is larger than 350. This lets the algorithm converge in just *two* iterations. The asymptotic rate of convergence is a whopping $m = 19$ (that of QDWH is cubic and scaled Newton is quadratic), although this is too high to observe numerically.

An important advantage of using $m = 2k+1$ with $k > 1$ is that each iteration is highly parallelizable. Specifically, each iteration requires computing the matrix rational function $r_j(X)$ of type $(m, m-1)$, which can be implemented using the partial fraction form by computing the following for $i =$

$1, \ldots, k$:

$$\begin{cases} \begin{bmatrix} \sqrt{c_i}X \\ I \end{bmatrix} &= \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R, \\ Y_i &= \frac{b_i}{c_i} X + \frac{1}{\sqrt{c_i}} \left(a_i - \frac{b_i}{c_i} \right) Q_1 Q_2^H \end{cases} \tag{1}$$

Here a_i, b_i, c_i are scalar parameters determined by the Zolotarev approximants. The k matrices Y_i can be computed completely independently from each other. Furthermore, as in QDWH, each Y_i is formed using a QR factorization and a matrix multiplication, both of which have been successfully implemented in parallel. Hence each iteration is highly parallelizable with two levels of parallelism.

The resulting algorithms for the polar, symmetric eigenvalue and singular value decompositions require slightly higher arithmetic cost than the QDWH-based algorithms but are better suited for parallel computing. Indeed, assuming that the computation of (1) is done independently for each i, the arithmetic cost along the critical path is much smaller than QDWH, and comparable with the standard algorithms based on reduction to tridiagonal or bidiagonal forms. Numerical experiments demonstrate its excellent numerical stability comparable to QDWH, and typically better than those of standard algorithms.

References

[1] N. I. Akhiezer. *Elements of the Theory of Elliptic Functions*, volume 79 of *Translations of Mathematical Monographs*. American Mathematical Society, 1990.

[2] R. Byers and H. Xu. A new scaling for Newton's iteration for the polar decomposition and its backward stability. *SIAM J. Matrix Anal. Appl.*, 30:822–843, 2008.

[3] Y. Nakatsukasa, Z. Bai, and F. Gygi. Optimizing Halley's iteration for computing the matrix polar decomposition. *SIAM J. Matrix Anal. Appl.*, 31(5):2700–2720, 2010.

[4] Y. Nakatsukasa and N. J. Higham. Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the SVD. *SIAM J. Sci. Comp*, 35(3):A1325–A1349, 2013.

Revisiting Greedy Ordering Heuristics for Sparse Matrix Factorizations

Esmond G. Ng and Barry W. Peyton

Abstract

It is well known that sparse matrix factorizations suffer fill. That is, some of the zero entries in a sparse matrix will become nonzero during factorization. To reduce factorization time and storage, it is important to arrange the computation so that the amount of fill is kept small. It is also well known that the amount of fill is affected in part by how the rows and columns of the sparse matrix are permuted (or ordered). For simplicity, we focus on the Cholesky factorization of a sparse symmetric positive definite matrix, in which case the fill depends solely on the sparsity of the given matrix and on the choice of permutation (or ordering).

We use the following notation. Let A be an $n \times n$ symmetric positive definite matrix and P be an $n \times n$ permutation matrix. We denote the factorization of PAP^T by LL^T, where L is lower triangular. We use L_{*k} to denote the k-th column of L. The number of nonzero entries in a vector v is represented by $\mathrm{nnz}(v)$.

It is well known that finding an ordering to minimize fill in sparse Cholesky factorization is an NP-complete problem [9]. Thus, we rely on heuristics. In this talk, we consider greedy heuristics.

The minimum degree (MD) algorithm, introduced by Tinney and Walker [7], is probably the most well-known greedy heuristic. It reduces fill by finding a permutation P so that $\mathrm{nnz}(L_{*k})$ is minimized locally at step k of the factorization. The MD algorithm is the symmetric variant of a method first proposed by Markowitz [4]. A great deal of work has been done on reducing the runtime of the MD algorithm.

In [7], Tinney and Walker also introduced another greedy heuristic, commonly known as the minimum local fill (MF) or minimum deficiency algorithm. In the MF algorithm, the permutation is chosen so that the number of zero entries in the reduced matrix that become nonzero is as small as possible at each step of the factorization. Interestingly enough, the MD algorithm minimizes locally the number of operations in computing the factorization (since the operation count at step k, in the right-looking formulation, is proportional to the square of $\mathrm{nnz}(L_{*k})$). On the other hand, the MF algorithm minimizes locally the amount of fill introduced into the matrix.

However, the MF algorithm has not been as popular as the MD algorithm. There are several reasons why this is the case. The metric $\mathrm{nnz}(L_{*k})$ is easy to obtain. On the other hand, to determine the amount of fill that will be introduced at each step, some sort of look-ahead is needed, which may be expensive. Consequently, the general belief has been that the MF algorithm is much more expensive than the MD algorithm. In [6], the authors said that "while many of the enhancements described above for minimum degree are applicable to minimum local fill (particularly supernodes), runtimes are still prohibitive", while in [5], it was reported that a true implementation of the MF algorithm was on the average slower by two orders of magnitude than one of the well-known implementations of the MD algorithm. Another reason for the lack of popularity of the MF algorithm is the belief that the MF orderings are often just marginally better than the MD orderings [1].

In an early version of [5], Ng and Raghavan reported that their MF orderings, on average, resulted in 9% less fill and 21% fewer operations than their MD orderings. Furthermore, Ng and Luce have recently solved the following open problem [3]: an ordering that minimizes fill does not necessarily minimize the number of operations, and vice versa. Thus, there is a need to revisit greedy ordering heuristics.

As noted above, the MF algorithm requires the determination of the amount of fill that will be introduced when a column is eliminated (known as the column's *deficiency*). Some of the deficiencies will change and need to be recomputed as the algorithm proceeds. However, recomputing the deficiency of a column from scratch is expensive.

In [8], Wing and Huang described an elegant way for *updating* the deficiencies rather than recomputing them from scratch. Their updating scheme was mentioned a few times in the circuit simulation literature, but it apparently was not widely used and it certainly was not adopted by the sparse matrix community. We will describe in this talk our recent work on the Wing-Huang updating scheme. In particular, we will show that the worst-case time complexity of the MF algorithm with Wing-Huang updates is the *same* as that of the MD algorithm. We will also demonstrate that techniques for reducing the runtime of the MD algorithm (such as quotient graph representations, mass eliminations, etc.) are equally applicable in the efficient implementation of the MF algorithm with Wing-Huang updates.

Results from our preliminary implementation of the MF algorithm with Wing-Huang updates are encouraging. Over a collection of 48 sparse matrices from the Florida Sparse Matrix Collection, our MF algorithm with Wing-Huang updates is just 4.9 times more expensive than the minimum degree algorithm with multiple eliminations (MMD) [2] on average. Our MF orderings, on the average, produce 17% less fill and require 31% fewer operations than the MMD algorithm. On one large test matrix (3dtube), MF produces 29% less fill and requires 55% fewer operations.

References

[1] Iain S. Duff, Albert M. Erisman, and John K. Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, New York, NY, 2nd edition, 1989.

[2] Joseph W.H. Liu. Modification of the minimum-degree algorithm by multiple elimination. *ACM Trans. Math. Software*, 11(2):141–153, 1985.

[3] Robert Luce and Esmond Ng. On the minimum flops problem in the sparse Cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 35(1):1–21, 2014.

[4] H.M. Markowitz. The elimination form of the inverse and its application to linear programming. *Management Science*, 3(3):255–269, 1957.

[5] Esmond G. Ng and Padma Raghavan. Performance of greedy ordering heuristics for sparse Cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 20(4):902–914, 1999.

[6] Edward Rothberg and Stanley C. Eisenstat. Node selection strategies for bottom-up sparse matrix orderings. *SIAM J. Matrix Anal. Appl.*, 19(3):682–695, 1998.

[7] W.F. Tinney and J.W. Walker. Direct solution of sparse network equations by optimally ordered triangular factorization. *Proc. IEEE*, 55(11):1801–1809, 1967.

[8] O. Wing and J. Huang. SCAP - a sparse matrix circuit analysis program. In *Proceedings - IEEE International Symposium on Circuits and Systems*, pages 213–215, 1975.

[9] Mihalis Yannakakis. Computing the minimum fill-in is NP-complete. *SIAM J. Alg. Disc. Meth.*, 2(1):77–79, 1981.

Conditioning and Preconditioning of the Weakly-Constrained Optimal State Estimation Problem

N.K. Nichols, A. El-Said, A.S. Lawless and R.J. Stappers

Abstract

Data assimilation is a technique for determining an 'optimal' estimate of the current and future states of a dynamical system from a prior estimate, or model forecast, together with observations of the system. Applications arise in very large environmental problems where the number of state variables is $O(10^7 - 10^8)$ and the number of observations is $O(10^4 - 10^6)$. The errors in the prior estimate and in the observations are assumed to be random with known distributions and the solution to the optimal estimation problem is taken to be the maximum a posteriori likelihood estimate. With the aid of Bayes Theorem, the problem reduces to a very large nonlinear least squares problem, subject to the dynamical system equations. The problem is treated in practice using an approximate Gauss-Newton iterative method, where a linearized least squares problem is solved at each step of the procedure by an 'inner' gradient iteration procedure.

The conditioning of the linearized least squares problem depends on the covariances of the errors in the states and in the observations. The problem is generally ill-conditioned and regularization of the problem is required. Previously we have examined how different components of the assimilation system influence the conditioning of the least squares problem in the case where the dynamical system equations are assumed to be exact, that is, the system equations form an equality constraint on the objective function. We have derived theoretical bounds on the condition number of the problem and have demonstrated how the conditioning is affected by the observation accuracy and by the specified length-scales in the error covariance of the prior estimate. A commonly used preconditioning technique, in which the prior states are transformed to uncorrelated variables, has also been analysed. We have shown that the problem becomes more ill-conditioned with increasingly dense and accurate observations (eg. Haben et al, Tellus, 2011).

Here we extend these results to a 'weak-constraint' variational formulation of the data assimilation problem, where it is assumed that the dynamical system equations contain random errors with known distributions and are not exact. Two different forms of the problem can be established. We investigate the conditioning of these two forms as a function of the covariances of the errors in the prior, the model states and the observations and as a function of the observational frequency. We study different preconditioning techniques for the two forms of the problem and determine theoretical bounds on the conditioning of the assimilation schemes. Numerical experiments are presented illustrating the results.

New Properties of Vector Spaces of (Quasi) Linearizations

Froilán M. Dopico, Yuji Nakatsukasa and Vanni Noferini

Abstract

A matrix polynomial P is structured if there are algebraic properties of its coefficients that induce some symmetries in the spectrum. Several such structures, such as symmetry or palindromicity, often arise in practice. In the last decade, the development of structure-preserving linearizations of a structured matrix polynomial P that are easy to construct given the coefficients of P has received a lot of attention in the literature. The interest in achieving this task stems from its applications to designing structure preserving algorithms for structured polynomial eigenvalue problems. For many structures, the problem was solved for odd degree matrix polynomials, but a complete solution when P has even degree is not currently available. Yet, some partial solutions exist. For instance, there are structured pencils that are linearizations if P is a structured regular matrix polynomial, but not if P is singular. In this context, obtaining more knowledge about the singular case is important both as a theoretical result and in order to assess the numerical reliability of these linearizations when they are applied to regular polynomials that are very close to singular ones.

In this talk we investigate some new properties of the pencils belonging to one of the most relevant families of structured pencils: the vector space of pencils $\mathbb{DL}(P)$, introduced by Mackey, Mackey, Mehl and Mehrmann in [2] and further investigated by Higham, Mackey, Mackey and Tisseur [1]. It is well known that pencils in $\mathbb{DL}(P)$ are linearizations of the regular polynomial P if and only if no eigenvalue of P is also a root of the associated scalar *ansatz polynomial* v. Moreover, they are never linearizations if P is singular. We investigate the properties of the pencils of $\mathbb{DL}(P)$ when *they are not linearizations of* P and show that, even in this case, they allow us to recover partial spectral information of regular matrix polynomials and more importantly, under some generic conditions, *the whole spectral information and all the minimal indices of singular matrix polynomials.*

More specifically, we study the Kronecker canonical form of a pencil $L \in \mathbb{DL}(P)$ in two important cases where the result is not trivial. When P is regular, we determine the Kronecker canonical form of L when there is a root of the ansatz scalar polynomial v which is also an eigenvalue of P: we find that some spectral information of P can still be extracted from L in this case. When P is singular, we assume that the spectra of P and v are disjoint and we prove that the partial multiplicities of P are the same as the partial multiplicities of L, and the minimal indices of P are just the ones of L after discarding a certain number of zeros.

Applications to both singular and near-singular structured polynomial eigenvalue problems will be discussed. In particular we emphasize that as a consequence of the results in this talk the pencils of $\mathbb{DL}(P)$ are, to the best of our knowledge, the first example of structured pencils that allow to recover all the spectral information and minimal indices of a symmetric or palindromic P when it is singular and has even degree.

References

[1] N. J. Higham, D. S. Mackey, N. Mackey and F. Tisseur, *Symmetric linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., 29 (2006), pp. 143–159.

[2] D. S. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Vector spaces of linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 971–1004.

Investigation of Crouzeix's Conjecture via Optimization

Anne Greenbaum, Adrian S. Lewis, Michael L. Overton and Lloyd N. Trefethen

Abstract

Crouzeix's conjecture is a fascinating open problem in matrix theory. We present a new approach to its investigation using optimization. Let p be a polynomial of any degree and let A be a square matrix of any order. Crouzeix's conjecture is the inequality

$$\|p(A)\| \le 2\|p\|_{W(A)}.$$

Here the left-hand side is the 2-norm of the matrix $p(A)$, while the norm on the right-hand side is the maximum of $|p(z)|$ over $z \in W(A)$, the field of values (or numerical range) of A. It is known that the conjecture holds if 2 is replaced by 11.08 (Crouzeix 2007). The conjecture is known to hold for certain restricted classes of polynomials p or matrices A, including the following cases, where m denotes the degree of p and n denotes the order of A: (a) $p(z) = z^m$ (power inequality, Berger and Pearcy 1966), (b) $n = 2$, (c) $W(A)$ is a disk (Badea 2004 based on work of von Neumann 1951 and Okubo and Ando 1975), (d) $n = 3$ and $A^3 = 0$ (Crouzeix 2012), (e) A is a diagonal scaling of a lower Jordan block with a perturbation in the top right corner (Choi and Greenbaum 2012), (f) A is diagonalizable with an eigenvector matrix having condition number less than or equal to 2 and (g) A is normal (in which case the constant 2 can be improved to 1).

Let us define the *Crouzeix ratio*

$$f(p, A) = \frac{\|p\|_{W(A)}}{\|p(A)\|}.$$

The conjecture states that $f(p, A)$ is bounded below by $1/2$ independently of m and n. Note that $f(p, A)$ is unbounded above because if p is the minimal polynomial of A then the denominator is zero but the numerator is not (unless A is just a scalar multiple of the identity). We can think of f as a real valued map on $\mathbb{C}^m \times \mathbb{C}^{n\times n}$ by associating p with its vector of coefficients $c \in \mathbb{C}^m$ using the monomial basis. The function f is nonconvex and nonsmooth (by the latter we mean it is not differentiable everywhere). It is not defined if $p(A) = 0$, but at all other points it is continuous. In fact, except at points where $p(A) = 0$, f is both locally Lipschitz and semialgebraic, because the norm in the numerator is attained on the boundary of $W(A)$ which can be parameterized by

$$z(\theta) = v(\theta)^* A v(\theta), \quad \theta \in [0, 2\pi)$$

where $v(\theta)$ is a normalized eigenvector corresponding to the largest eigenvalue of the Hermitian matrix

$$H = \frac{1}{2}(e^{i\theta}A + e^{-i\theta}A^*).$$

If $W(A)$ is just a line segment, we interpret the boundary to be the whole set.

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is a remarkably effective method for minimizing nonsmooth, nonconvex functions and in practice it reliably generates sequences converging to locally minimal values, although its convergence theory is quite limited. The ratio f can be computed for a given p and A using the Chebfun toolbox, which provides an extremely convenient way to compute with functions that are represented to machine precision accuracy using Chebshev polynomial interpolation. In fact, `fov` is a built-in Chebfun function for computing the boundary of $W(A)$, exploiting its characterization given above. This allows us to compute the Crouzeix ratio f in a single line of MATLAB:

```
max(abs(polyval(c, fov(A))))/norm(polyvalm(c, A))
```

Here `c` is the vector of p's coefficients while `max, abs` and `polyval` are overloaded versions of standard MATLAB functions that are applicable to chebfuns. The main cost is the construction of the chebfun defining the field of values. The BFGS method also requires the gradient of f with respect to c and A; deriving the formulas is a rather lengthy exercise ("the chain rule on steroids") but the resulting code provides the gradient for essentially no extra cost.

We will report on extensive experiments searching for minima of f for fixed m and n. So far, our experiments strongly support the truth of Crouzeix' conjecture. We have found it to be especially interesting to optimize over A while keeping p fixed, which has led to questions such as the following.

Questions. Let p with real coefficients and degree m be fixed.

1. Under what conditions on p does there exist a real matrix A with $f(p, A) = 1/2$? The only cases we know are $p(z) = (z-\gamma)^m$ for some $\gamma \in \mathbf{R}$ and $m = 1$ or $m = 2$. Then, set $n = m+1$ and $A = \gamma I + N$, where N has nonzeros only on its first subdiagonal, each of whose entries is set to $\pm\xi$, for some $\xi \neq 0$: then $p(A) = \pm\xi^m e_n e_1^T$, $W(A)$ is a disk centered at γ with radius $|\xi|\cos(\pi/(n+1)) = |\xi|/2^{1/m}$ and so $\|p\|_{W(A)} = |\xi|^m/2 = \|p(A)\|/2$.

2. Is it the case that the value $1/2$ for the ratio f can be arbitrarily well approximated by a real bounded sequence of matrices A if and only if p has a real root? Suppose $m = 2$ and p has two real roots γ and β: set $n = 2$ and $A = \gamma I + \xi e_2 e_1^T$ or $\beta I + \xi e_2 e_1^T$. Then $p(A) = \pm(\gamma-\beta)\xi e_2 e_1^T$, $W(A)$ is a disk centered at γ or β with radius $|\xi|/2$ and so

$$\|p\|_{W(A)} = \frac{|\xi|}{2}\left(|\gamma-\beta| + \frac{|\xi|}{2}\right) = \left(\frac{1}{2} + \frac{|\xi|}{4|\gamma-\beta|}\right)\|p(A)\|,$$

   so the ratio converges to $1/2$ as $\xi \to 0$. The value $1/2$ is not attained since when $\xi = 0$ both the numerator and denominator of f are 0. Furthermore, if we consider the more general sequence $f(p, \gamma I + \Delta A)$ as $\Delta A \to 0$, any limiting value in $[1/2, 1]$ can be approximated by suitable choice of the sequence ΔA.

3. If the answer to the previous question is yes, is it the case that for any fixed n, and all ϕ sufficiently close to, but greater than, $1/2$, each real root of p corresponds to a different nonempty connected component of the sublevel set of matrices $\{A \in \mathbb{R}^{n\times n} : f(p, A) \leq \phi\}$?

4. Diagonal real matrices, whose fields of values are line segments, are stationary points of $f(p, \cdot)$ in $\mathbb{R}^{n\times n}$, that is, $\nabla_A f$ is zero at diagonal matrices, but they are not necessarily local minimizers. Is it the case that if p has only real roots, then $1/2$ and 1 are the only stationary or locally minimal values of f? If p has complex roots do there always exist local minima with values strictly between 0.5 and 1?

5. Finally, let p be fixed with complex coefficients and degree m. When does there exist a complex matrix A such that $f(p, A) = 1/2$? Is it the case that it is always possible to find a bounded sequence of complex matrices A so that $f(p, A)$ approximates $1/2$ to any accuracy?

The ultimate goal is to prove Crouzeix's conjecture. Whether the experiments and resulting new questions will help remains to be seen, but we will report on the latest developments in this direction.

Augmented error analyses of Vector Orthogonalization and related Algorithms

Chris Paige, Ivo Panayotov, Wolfgang Wülling and Jens-Peter Zemke

Abstract

Any matrix $V_k = [v_1, \ldots, v_k]$ of k unit 2-norm n-vectors can be augmented to give $(n+k) \times k$ $Q_1^{(k)} = \begin{bmatrix} S_k \\ V_k(I-S_k) \end{bmatrix}$ such that $Q_1^{(k)H} Q_1^{(k)} = I_k$. Here $S_k = (I + U_k)^{-1} U_k$ where U_k is the strictly upper triangular part of $V_k^T V_k = U_k^T + I_k + U_k$. Modified Gram Schmidt (MGS) for a given $n \times k$ matrix A ideally produces $n \times k$ V_k and $k \times k$ upper triangular R_k such that $A = V_k R_k$, $V_k^T V_k = I_k$. In 1992 Björck & Paige used the above augmentation to show that finite precision MGS leads to a backward stable augmented system $\begin{bmatrix} E_1 \\ A+E_2 \end{bmatrix} = Q_1^{(k)} R_k$, $Q_1^{(k)}$ as above, $\|E_i\|_2 \le O(\epsilon)\|A\|_2$.

This approach showed that we obtain backward stable solutions of linear least squares and some related problems using MGS (Björck & Paige, 1992, 1994), and of $Ax = b$ using MGS-GMRES (Paige, Rozložník, & Strakoš, 2006). Later this approach was used in [1] to show that the Lanczos process for producing a symmetric tridiagonal matrix T_k from a symmetric matrix A, which ideally gives $AV_k = V_{k+1} T_{k+1,k}$, $V_{k+1}^T V_{k+1} = I$, computationally satisfies a strange backward stable Lanczos process. That is, with $Q_1^{(k)}$ above, the computed $T_{k+1,k}$ satisfies

$$\left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H^{(k)} \right) Q_1^{(k)} = [Q_1^{(k)}, q_{k+1}]\, T_{k+1,k}, \quad H^{(k)} = H^{(k)T}, \quad \|H^{(k)}\|_2 \le O(\epsilon)\|A\|_2.$$

Let $\widetilde{Q}_1^{(k)}$ be $Q_1^{(k)}$ less its zero k-th row. Then this formulation will, by eliminating the kth column of $\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix}$ (and of $H^{(k)}$ to give $\widetilde{H}^{(k)}$) and the k-th row of $Q_1^{(k)}$, be rewritten below to show how

$$\begin{array}{ccccccc} T_{k,k-1} & \to & T_{k+1,k} & \& & \widetilde{Q}_1^{(k)} & \to & \widetilde{Q}_1^{(k+1)} \\ k \times (k-1) & & (k+1) \times k & & (n+k-1) \times k & & (n+k) \times (k+1). \end{array}$$

$$\left(\begin{bmatrix} T_{k,k-1} & 0 \\ 0 & A \end{bmatrix} + \widetilde{H}^{(k)} \right) \widetilde{Q}_1^{(k)} = \widetilde{Q}_1^{(k+1)} T_{k+1,k}, \quad \|\widetilde{H}^{(k)}\|_2 \le O(\epsilon)\|A\|_2, \tag{1}$$

$$\widetilde{Q}_1^{(k)} = \begin{bmatrix} [S_{k-1}, s_k] \\ V_k(I-S_k) \end{bmatrix}, \quad \text{where} \quad S_k = \begin{bmatrix} S_{k-1} & s_k \\ 0 & 0 \end{bmatrix},$$

$$\widetilde{Q}_1^{(k+1)T} \widetilde{Q}_1^{(k+1)} = [Q_1^{(k)} \mid q_{k+1}]^T [Q_1^{(k)} \mid q_{k+1}] = I_{k+1}.$$

This shows that the new $\{k,k\}$ and $\{k+1,k\}$ elements α_k and β_{k+1} of the tridiagonal matrix $T_{k+1,k}$ depend on the whole of $T_{k,k-1}$ as well as on A. The above augmented Lanczos process indicates why the computational process can produce T_k having many repeats of eigenvalues of A.

These algorithms can be called vector orthogonalization algorithms, and in some of these, the supposedly orthonormal vectors $\tilde{v}_j$ produced by the finite precision processes can quickly lose orthogonality, but are all nearly unit 2-norm vectors. When each $\tilde{v}_j$ is normalized to the corresponding v_j having unit norm, this gives $V_k = [v_1, \ldots, v_k]$ above. It can be shown that $\|S_k\|_2 \le 1$, and the number of unit singular values of S_k is *exactly* the rank deficiency of V_k, and so of $\widetilde{V}_k \equiv [\tilde{v}_1, \ldots, \tilde{v}_k]$, see [1, Corollary 2.2]. Thus S_k contains all we need to know about the loss of orthogonality and rank of $\widetilde{V}_k$. The above augmented systems also contain S_k, and so not only describe the development of the desired computed elements, but the development of the loss of orthogonality matrix S_k.

While this approach has led to successful rounding error analyses of the underlying methods based on MGS which orthogonalize against all previous vectors, the analysis for methods like the Lanczos process based on *implicit* orthogonalization appears to be far more complicated, as (1) suggests. This indicates that a deeper understanding of the orthogonal matrix $Q_1^{(k)}$ is required, see [3].

Such analyses could help in understanding all vector orthogonalization algorithms, and we illustrate some of the insights so far gained. First we will give a simpler development of these augmented results than previously—a development we suspect is quite general. Then we will show that the approach can also be applied to bi-orthogonalization processes, see [2]. Finally we will discuss some properties of the orthogonal matrix $Q_1^{(k)}$ given in [3], along with our more recent findings.

References

[1] C. C. Paige, *An Augmented Stability Result for the Lanczos Hermitian Matrix Tridiagonalization Process*. SIAM. J. Matrix Anal. Appl., 31, Issue 5 (2010), pp. 2347-2359.

[2] C. C. Paige, I. Panayotov, and J.-P. M. Zemke, *An Augmented Analysis of the Perturbed Two-sided Lanczos Tridiagonalization Process*. Linear Algebra Appl., accepted May 2013.

[3] C. C. Paige and W. Wülling *Properties of a unitary matrix obtained from a sequence of normalized vectors.* Submitted to SIAM J. Matrix Anal. Appl., November 2012.

The Fiedler Companion Matrix

Beresford N. Parlett

Abstract

This matrix has the same entries as the standard companion matrix for a given polynomial but arranged in a special way as a pentadiagonal matrix. It, and its LR transforms, have properties in common with several different classes: banded matrices with banded inverses (G. Strang), CMV matrices, and quasiseparable matrices. It is easy to reduce the Fiedler matrix (balanced) to upper Hessenberg form and it is then a rival to the traditional companion matrix as input to the QR algorithm to determine its spectrum. However the LR and qd transforms preserve the original nonzero structure and so reduce the arithmetic effort compared to QR.

The Development of Preconditioned Iterative Solvers for PDE-Constrained Optimization Problems

John W. Pearson, Martin Stoll and Andrew J. Wathen

Abstract

In past decades one of the main application areas of numerical linear algebra has been in developing iterative methods to accurately solve partial differential equations (PDEs). Typically such iterative methods incorporate preconditioning strategies to accelerate the convergence of these schemes. Such approaches are now very well understood for a vast number of PDEs, and have been utilized to tackle a wide range of industrial applications.

Of late, a related field that has been widely considered in scientific computing and applied sciences is that of PDE-constrained optimization, that is a problem where a functional is sought to be minimized with one or more PDEs acting as constraints. Such problems have applications in many areas, including flow control, medical imaging, option pricing, biological and chemical processes and electromagnetic inverse problems, to name but a few. Often, solving these problems numerically (in our case using a finite element method) involves solving a large and sparse matrix system that is of saddle point structure – the preconditioning strategies developed for forward PDEs therefore have considerable applicability to related PDE-constrained optimization problems.

In this presentation, we outline a framework for devising iterative methods for a range of matrix systems resulting from PDE-constrained optimization, which are of various structures, and arise from different application areas. We approach the solution of these matrix systems by exploiting their saddle point structure, and creating effective approximations of the $(1,1)$-block and Schur complement of the matrices involved. We employ many different techniques to achieve this, including mass matrix approximation, various Schur complement factorizations, commutator arguments to approximate matrices arising in problems from fluid dynamics, and techniques to deal with time-dependent and nonlinear terms within the problem statement. If good approximations may be obtained using approaches such as these, then it is possible to create powerful iterative solvers based around Krylov subspace methods such as MINRES, non-standard Conjugate Gradients, GMRES and BiCG. The aim when creating our solvers is that they are feasible, reliable, and robust with respect to the dimension of the matrix system as well as the parameters involved in the problem.

We discover that our strategies lead to methods with these properties for a wide range of PDE-constrained optimization problems. We discuss a number of such problems within this presentation, commencing with the fundamental Poisson control [1, 2] and convection-diffusion control [3] problems, before extending our methods for these simpler formulations to develop strategies for Stokes [4] and Navier-Stokes control setups [5], which are motivated by fluid dynamic processes. We also discuss how these methods may be adapted to solve time-dependent PDE-constrained optimization problems [6], for which the matrix systems are of very high dimension even in comparison to those for related time-independent problems. Additionally, we examine reaction-diffusion control problems resulting from the modelling of chemical processes [7], which include both time-dependent and nonlinear terms within the problem setup. For each problem we examine, we motivate the preconditioning strategies used, stating eigenvalue bounds for the preconditioned systems where relevant, and present results from numerical experiments to demonstrate the potency of our approaches.

We also aim to provide an outlook of the field, and highlight possible areas of future research into preconditioned iterative methods for matrix systems arising from PDE-constrained optimization. In particular, we discuss the potential for parallelization of a number of the methods described

here, and present further applications which could be tackled by strategies similar to those applied in this talk.

References

[1] J. W. Pearson, and A. J. Wathen, *A New Approximation of the Schur Complement in Preconditioners for PDE-Constrained Optimization*, Numerical Linear Algebra with Applications, 19(5), pp.816–829, 2012.

[2] J. W. Pearson, M. Stoll, and A. J. Wathen, *Preconditioners for State Constrained Optimal Control Problems with Moreau-Yosida Penalty Function*, to appear in Numerical Linear Algebra with Applications, available online, DOI: 10.1002/nla.1863, 2012.

[3] J. W. Pearson, and A. J. Wathen, *Fast Iterative Solvers for Convection-Diffusion Control Problems*, Electronic Transactions on Numerical Analysis, 40, pp.294–310, 2013.

[4] J. W. Pearson, *On the Role of Commutator Arguments in the Development of Parameter-Robust Preconditioners for Stokes Control Problems*, submitted to Electronic Transactions on Numerical Analysis, 2013.

[5] J. W. Pearson, *Preconditioned Iterative Methods for Navier-Stokes Control Problems*, submitted to SIAM Journal on Scientific Computing, 2013.

[6] J. W. Pearson, M. Stoll and A. J. Wathen, *Regularization-Robust Preconditioners for Time-Dependent PDE-Constrained Optimization Problems*, SIAM Journal on Matrix Analysis and Applications, 33(4), pp.1126–1152, 2012.

[7] J. W. Pearson, and M. Stoll, *Fast Iterative Solution of Reaction-Diffusion Control Problems Arising from Chemical Processes*, SIAM Journal on Scientific Computing, 35, pp.B987–B1009, 2013.

GMRES Convergence Bounds that Depend on the Right-Hand Side Vector

David Titley-Peloquin, Jennifer Pestana and Andrew Wathen

Abstract

GMRES [7] remains one of the most popular methods for solving large, sparse linear systems $Ax = b$, $A \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$. The k-th GMRES iterate x_k, with corresponding residual vector $r_k \equiv b - Ax_k$, is defined by

$$\|r_k\|_2 = \min_{\substack{q\in\Pi_k \\ q(0)=1}} \|q(A)r_0\|_2, \tag{1}$$

where Π_k denotes the set of polynomials of degree at most k.

Despite the widespread use of GMRES, obtaining generally descriptive convergence bounds remains a difficult problem. A typical first step is to separate A and r_0 to obtain the ideal GMRES problem [2]

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \min_{\substack{q\in\Pi_k \\ q(0)=1}} \|q(A)\|_2. \tag{2}$$

For a diagonalizable A, with diagonalization $A = Z\Lambda Z^{-1}$, $\Lambda = \mathrm{diag}(\lambda_i)$, we then arrive at the textbook bound

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \|Z\|_2\|Z^{-1}\|_2 \min_{\substack{q\in\Pi_k \\ q(0)=1}} \|q(\Lambda)\|_2. \tag{3}$$

Although simpler, (2) and (3) are worst-case bounds that hold for every $r_0 \in \mathbb{C}^n$. Usually, however, one has a specific r_0 that may be far from the worst case and for which (2) and (3) may not be descriptive. To rectify this, bounds that explicitly include the right-hand side vector have been considered for specific examples such as Jordan blocks [3] and tridiagonal Toeplitz matrices [4, 6], as well as for more general matrices [5].

In this talk, which is based on work in [9], we present simple right-hand side dependent bounds for GMRES applied to a linear system with diagonalizable matrix A. The bounds capture interactions between r_0 and A through $w = Z^{-1}r_0/\|r_0\|_2$, a vector that contains the co-ordinates of r_0 in the eigenvector basis. The first formulates convergence in terms of the least-squares approximation

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \|Z\|_2 \min_{\substack{q\in\Pi_k \\ q(0)=1}} \|Wq(\Lambda)e\|_2, \tag{4}$$

where $W = \mathrm{diag}(w_i)$ and $e = [1, \ldots, 1]^T$. It shows that the GMRES relative residual is bounded above by the product of $\|Z\|_2$ and the residual of a weighted polynomial least-squares approximation problem on the spectrum of A. Further, all nonnormality can be fully contained in the weight matrix W. We characterize situations in which (4) is descriptive and prove that it can never be worse than (3).

A second GMRES bound, that is based on an ideal GMRES problem for a rank-1 perturbation of the diagonal matrix of eigenvalues Λ, will also be discussed. The perturbation involves both the eigenvalues and w, so that again the right-hand side, and its interaction with A, are incorporated.

Numerical experiments show that our bounds can be quantitatively descriptive of GMRES convergence for unpreconditioned and preconditioned problems, including those for which bounds based

on the standard ideal GMRES problem (2) fail. As a particular example, we show that our bounds can be applied to ill-posed problems and that in this case they, similarly to the analysis in [1], explain the semi-convergence typically observed.

We then consider how to modify the linear system to reduce problematic nonnormality. We apply an unusual form of "preconditioning" that preserves eigenvalues but that modifies both the right-hand side vector and the eigenvectors to reduce the size of components of w.

References

[1] L. Eldén and V. Simoncini. Solving ill-posed linear systems with GMRES and a singular preconditioner. *SIAM J. Matrix Anal. Appl.*, 33, 1369–1394, 2012.

[2] A. Greenbaum and L. N. Trefethen. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM J. Sci. Comput.*, 15, 359–368, 1994.

[3] I. C. F. Ipsen. Expressions and bounds for the GMRES residual. *BIT*, 40, 524–535, 2000.

[4] R.-C. Li and W. Zhang. The rate of convergence of GMRES on a tridiagonal Toeplitz linear system. *Numer. Math.*, 112, 267–293, 2009.

[5] J. Liesen. Computable convergence bounds for GMRES. *SIAM J. Matrix Anal. Appl.*, 21, 882–903, 2000.

[6] J. Liesen and Z. Strakoš. Convergence of GMRES for tridiagonal Toeplitz matrices. *SIAM J. Matrix Anal. Appl.*, 26, 233–251, 2004.

[7] Y. Saad, and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7, 856–869, 1986.

[8] V. Simoncini and D. B. Szyld. GMRES: Recent computational developments in Krylov subspace methods for linear systems. *Numer. Linear Algebra Appl.*, 14, 1–59, 2007.

[9] D. Titley-Peloquin, J. Pestana and A. J. Wathen. GMRES convergence bounds that depend on the right-hand-side vector. *IMA J. Numer. Anal.*, Published online, 2013.

The Core Problem within a Linear Approximation Problem with Multiple Right-Hand Sides

Iveta Hnětynková, Martin Plešinger and Diana M. Sima

Abstract

Consider a linear (orthogonally invariant) approximation problem with multiple right-hand sides

$$AX \approx B, \qquad \text{where} \qquad A \in \mathbb{R}^{m\times n}, \quad X \in \mathbb{R}^{n\times d}, \quad B \in \mathbb{R}^{m\times d}.$$

The *total least squares* (TLS) formulation seeks for a solution X of

$$(A+E)X = B+G \qquad \text{such that} \qquad \min \|[G,E]\|_F.$$

A question of existence and uniqueness of this *TLS solution* has been studied for decades. Golub and Van Loan in the paper [1] showed that even with $d = 1$ the TLS solution may not exist, and when it exists, it may not be unique. The book [5] by Van Huffel and Vandewalle introduced the *nongeneric* approach, extended the Golub–Van Loan's analysis to *two special cases* with $d \geq 1$, and gave the so-called *classical TLS algorithm*. Wei further analyzed problems with nonunique solutions in [6, 7]. At Householder Symposium XVIII, 2011, we presented necessary and sufficient condition for existence of TLS solution in the *general case* (with $d \geq 1$). Our analysis, based on [1, 5], resulted in a new classification of TLS problems; see [2].

The single right-hand side case ($d = 1$) was revisited by Paige and Strakoš in [4]. They introduced a minimally dimensioned subproblem of $Ax \approx b$ called a *core problem* always having the unique TLS solution. At Householder Symposium XVIII, we extended the core problem concept to the general case $d \geq 1$. Definition and detailed analysis of this core problem can be found in the recent paper [3].

In this contribution we concentrate on solvability of the core problem for $d > 1$. Using the properties of the right singular vector subspaces of the corresponding extended core problem matrix $[B_1, A_{11}]$, it will be shown that the core problem with multiple right-hand sides *may not have a TLS solution*. We show that core problems with multiple right-hand sides can have internal structure which allows to interpret the original problem as a direct sum of two (or more) *uncorrelated components*. In such case we call the core problem *reducible*. It will be shown that existence of a TLS solution of a reducible core problem depends on existence of a TLS solution of its components, but also on the relations among singular values of these components. Finally we show that also an *irreducible* core problem with multiple right-hand sides may not have a TLS solution.

References

[1] Golub, G. H., Van Loan, C. F.: An analysis of the total least squares problem, Numer. Anal. **17** (1980), pp. 883–893.

[2] Hnětynková, I., Plešinger, M., Sima, D. M., Strakoš, Z., Van Huffel, S.: The total least squares problem in $AX \approx B$. A new classification with the relationship to the classical works, SIAM J. Matrix Anal. Appl. **32** (2011), pp. 748–770.

[3] Hnětynková, I., Plešinger, M., Strakoš, Z.: The core problem within a linear approximation problems $AX \approx B$ with multiple right-hand sides, SIAM J. Matrix Anal. Appl. **34** (2013), pp. 917–931.

[4] Paige, C. C., Strakoš, Z.: Core problem in linear algebraic systems, SIAM J. Matrix Anal. Appl. **27** (2006), pp. 861–875.

[5] Van Huffel, S., Vandewalle, J.: The Total Least Squares Problem: Computational Aspects and Analysis, SIAM Publications, Philadelphia, PA, 1991.

[6] Wei, M.: The analysis for the total least squares problem with more than one solution, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 746–763.

[7] Wei, M.: Algebraic relations between the total least squares and least squares problems with more than one solution, Numer. Math., 62 (1992), pp. 123–148.

Computing all Values λ such that $A + \lambda B$ has a Multiple Eigenvalue

Andrej Muhič and Bor Plestenjak

Abstract

For a given pair of $n \times n$ complex matrices A and B we want to compute all values λ such that the matrix $A + \lambda B$ has a multiple eigenvalue. We give construction of matrices Δ_1 and Δ_0 of size $3n^2 \times 3n^2$, such that the regular finite eigenvalues of the singular pencil $\Delta_1 - \lambda\Delta_0$ are exactly the values λ, such that $A + \lambda B$ has a multiple eigenvalue. This provides new insight into the problem and allows us to compute the values numerically from Δ_1 and Δ_0 by the staircase algorithm, using standard eigenvalue computation tools.

The problem was introduced in 2011 by E. Jarlebring, S. Kvaal, and W. Michiels. They provide a method of fixed relative distance (MFRD), which works as follows. If $A + \lambda_0 B$ has a double eigenvalue, then, for λ close to λ_0, the matrix $A + \lambda B$ has eigenvalues μ and $(1 + \epsilon)\mu$, where ϵ is small. We can write this as a two-parameter eigenvalue problem

$$\begin{aligned} (A + \lambda B - \mu I)x &= 0 \\ (A + \lambda B - \mu(1 + \epsilon)I)y &= 0, \end{aligned}$$

which is nonsingular for a given $\epsilon \neq 0$. The solutions of the above two-parameter eigenvalue problem are used as initial approximations for Newton's method that computes the final solutions.

We show that the $3n^2 \times 3n^2$ matrices

$$\begin{aligned} \Delta_0 &= B \otimes R + I \otimes Q \\ \Delta_1 &= -I \otimes P - A \otimes R, \end{aligned}$$

where

$$P = \begin{bmatrix} A^2 & AB + BA & -2A \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & B^2 & -B \\ -I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & -B & I \\ 0 & 0 & 0 \\ -I & 0 & 0 \end{bmatrix}$$

form a singular pencil $\Delta_1 - \lambda\Delta_0$ with a property that its finite regular eigenvalues agree with the values λ, such that $A + \lambda B$ has a multiple eigenvalue. In particular, for the generic solution, when μ_0 is a double eigenvalue of $A + \lambda_0 B$, we show that

a) a semisimple μ_0 contributes four linearly independent vectors to the null space and the rank of the pencil $\Delta_1 - \lambda\Delta_0$ drops by two at $\lambda = \lambda_0$,

b) a non-semisimple μ_0 contributes three linearly independent vectors to the null space and the rank of the pencil $\Delta_1 - \lambda\Delta_0$ drops by one at $\lambda = \lambda_0$.

This is an example of a problem where one has to compute all or some of the finite regular eigenvalues of a singular matrix pencil. The staircase algorithm does the job, but, due to its large sensitivity, can only be applied to small matrix pencils. When dealing with singular matrix pencils, one is usually satisfied with modest approximations of the finite regular eigenvalues. We will present some heuristics that could be used for this task.

Triplet Representations for Solving Matrix Equations in Queuing Theory

Giang T. Nguyen and Federico Poloni

Abstract

Let $Z \in \mathbb{R}^{n\times n}$ be a matrix with $Z_{ij} \leq 0$ whenever $i \neq j$. A *triplet representation* for Z is a triple $(\operatorname{offdiag}(Z), u, v) \in \mathbb{R}^{n^2-n}_{\leqslant 0} \times \mathbb{R}^n_{>0} \times \mathbb{R}^n_{\geqslant 0}$, where $\operatorname{offdiag}(Z) \leqslant 0$ is a vector containing (in some specified order) the offdiagonal elements of Z, and $u > 0$ and $v \geqslant 0$ are vectors such that $Zu = v$. Here, inequalities are intended in the componentwise sense.

Matrices for which a triplet representation exists are known as *M-matrices*, and several classical alternative characterizations for them exist. A triplet representation allows one to uniquely reconstruct Z.

For M-matrices, there exists a variant of Gaussian elimination that can be computed starting from a triplet representation and is *subtraction-free*, i.e., it requires multiplications, divisions and additions of nonnegative numbers only [2]. In particular, one can bound the rounding error for each of these operations with the machine precision. Hence several operations with M-matrices, such as solving linear systems with a nonnegative right-hand side $b \geq 0$ or computing one-dimensional kernels, are perfectly forward stable (irrespective of their condition number) when one has a triplet representation available. Moreover, they are stable with respect to *componentwise error*

$$\operatorname{cwerr}(\hat{X}, X) = \max\left(\left|\hat{X} - X\right| ./ |X|\right),$$

where ./ denotes component-by-component division and we agree that $0./0 = 0$.

The catch here is that computing a triplet representation from the entries of the matrix itself is ill-conditioned. As an example, consider $Z = \begin{bmatrix} 1 & -1 \\ -1 & 1+\varepsilon \end{bmatrix}$: its LU factorization is $\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 0 & \varepsilon \end{bmatrix}$, and for it to be correct Z_{22} needs to be known with high accuracy. However, the triplet representation $\left(\operatorname{offdiag}(M) = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, v = \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix}\right)$ is less sensitive to relative perturbations of each entry since it contains ε explicitly, and allows to compute the factorization more accurately.

For several applications in applied probability, triplet representations are known *a priori* and are crucial to guarantee high componentwise accuracy. This kind of error bounds is particularly welcome there, since the probabilities of interest are quantities that may span over many orders of magnitudes.

We focus here on solving matrix equations from so-called *fluid queue* models [5]. These are probabilistic models for the expected quantity of "fluid" in a "bucket", where the in- and out-flow vary according to the state of a continuous time Markov chain (modelling the environment). They are popular in modelling computer networks and buffers, and are formally similar to some boundary-value problems appearing in control theory, although the involved matrix structures are completely different. Computing the solution of interest is equivalent to finding a special invariant subspace of the matrix polynomial

$$\lambda^2 V - \lambda D + Q^T = 0, \tag{1}$$

where Q is the generator of the Markov chain (hence $-Q$ is a singular M-matrix), $V \geqslant 0$ is diagonal with nonnegative entries, and D is diagonal with mixed signs.

The case in which $V = 0$ (first-order models) has been studied more extensively; it is equivalent to solving a nonsymmetric algebraic Riccati equation with special sign structure. The paper [4] provides a detailed componentwise error analysis and suggests using triplet representations to devise a componentwise stable method based on the structured doubling algorithm (SDA). Several matrices have to be inverted along the algorithm; the suggestion there is computing triplet representations for them from scratch at each step. This is potentially dangerous, as triplet representations are an ill-conditioned function of the matrix entries whenever the matrix is close to singular. We show here that the triplet representation can be updated along the algorithm instead, together with the other matrices, with subtraction-free formulas, and thus there is no need to recompute them. This modification to the algorithm leads to a significant increase in stability with respect to the available methods, as shown by our numerical experiments. The ideas used to obtain these triplet representations are related to a novel probabilistic interpretation of SDA for this model: the computed quantities can be interpreted explicitly as transition probabilities in a discretized version of the problem.

The quadratic case with $V \neq 0$ (second-order models) has not been considered in similar detail in literature, up to our knowledge. Algorithms for solving it have been proposed, but they are based either on computing eigenvectors explicitly or on generalized Schur decompositions [1]; thus they are backward stable in the normwise sense, but the smallest entries of these vectors will typically be computed with poor relative accuracy.

We focus here on Cyclic Reduction (CR), a classical algorithm for several problems that are related to quadratic matrix polynomials and quadratic matrix equations. Cyclic Reduction has been applied to several probabilistic problems; under suitable conditions the matrices appearing along the algorithm keep the same sign structure (some are nonnegative, some are M-matrices). Its use has been suggested recently [3] to solve models associated with (1), too, but the sign structures are not preserved in this case and it is not apparent how to obtain a subtraction-free variant.

We show that one can apply a suitable nonlinear transformation to the associated matrix polynomial (1) to correct the signs and obtain a subtraction-free version of CR for this problem as well. This transformation is related to shifting techniques, and is an extension to the quadratic case of the known relations between CR and SDA. Together with this formulation that preserves the sign structure, we show that explicit (subtraction-free) formulas to update triplet representations are available for Cyclic Reduction, too; hence the whole computation can be carried on with high componentwise accuracy guarantees.

References

[1] M. Agapie, K. Sohraby. Algorithmic solution to second-order fluid flow. IEEE Infocom 2001.

[2] A. S. Alfa, J. Xue, Q. Ye. Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comp. 71 (2002).

[3] G. Latouche, G. T. Nguyen. The morphing of fluid queues into Markov-modulated Brownian motion. Submitted.

[4] W.-G. Wang, W.-C. Wang, R.-C. Li. Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations. SIAM J. Matrix Anal. Appl. 33 (2012).

[5] V. Ramaswami, Matrix analytic methods for stochastic fluid flows, in: D. Smith, P. Hey

(Eds.), Teletraffic Engineering in a Competitive World (Proceedings of the 16th International Teletraffic Congress), 1999.

Efficient Computation of the Posterior Covariance Matrix in Large-Scale Variational Data Assimilation Problems

Kirsty Brown, Igor Gejadze and Alison Ramage

Abstract

For a large-scale variational data assimilation problem, the Posterior Covariance Matrix (PCM) can be used to obtain important information, such as confidence intervals for the optimal solution. However, due to memory limitations, in practice it is often impossible to assemble, store or manipulate the PCM explicitly for a realistic model. One alternative approach is to approximate the PCM by the inverse Hessian of the auxiliary control problem based on the tangent linear model constraints. In many practical data assimilation problems, the majority of eigenvalues of the inverse Hessian are clustered in a narrow band above the unity, with only a relatively small percentage of the eigenvalues being distinct from this. A limited-memory representation of the inverse Hessian can therefore be built on the basis of a small (compared to the dimension of the state vector) number of 'leading' Ritz pairs of the projected Hessian, computed by the Lanczos method.

Although this approach is useful, taking into account the size of the state vectors used in modern realistic data assimilation applications (for example, 10^9 to 10^{12} unknowns), using even a small percentage of the spectrum could still involve a significant number of eigenvalues, which may be far beyond the available storage capacity. In such cases, it is usually necessary to enhance the limited-memory algorithm by using an appropriate preconditioner. Within this framework, it is standard to apply first-level preconditioning in the form of the square root of the background covariance matrix. However, if the impact of sensor information on the optimal solution is significant, this is not sufficient and an additional preconditioning step is required. In this talk, we will discuss using a multilevel approach as a second-level preconditioner and investigate its usefulness in terms of reducing memory requirements and increasing computational efficiency,

Preconditioning Linear Systems arising in Constrained Optimization Problems

Tyrone Rees

Abstract

One of the major bottlenecks in modern optimization codes is the requirement to solve multiple so-called saddle point systems of the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

For example, primal-dual interior point methods apply variants of Newton's method to a non-linear system to find a minimum of constrained quadratic program. This method requires the solution of a sequence of saddle point systems of the form above – called the augmented system in this context – at each step of the Newton iteration.

In many situations – especially for large scale optimization problems – we would like to solve the saddle point system iteratively, as computing the factorizations at each Newton step to employ a direct solver would be prohibitively expensive. However, it is important to choose the iterative solver so that the errors are sympathetic to the outer Newton iteration. For this reason Krylov subspace methods with constraint preconditioners, i.e., preconditioners of the form

$$P_{con} = \begin{pmatrix} G & B^T \\ B & 0 \end{pmatrix},$$

have been popular in the optimization community, as they are known to preserve the constraints, and this fact can be used to prove convergence of the inexact Newton method. However, applying such a preconditioner can be costly, and ensuring that the solution remains on the constraint manifold in the presence of rounding errors can be delicate.

Block diagonal preconditioners of the form

$$P_{bd} = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}$$

have proved successful in, e.g., the field of computational fluid dynamics, due to their ease of application and effectiveness. As the solution moves away from the constraint manifold, such preconditioners fall outside of the scope of the known theory of the convergence of the outer (Newton) iteration. However, I will describe how – as the outer iteration progresses – the iterates returned by Schur complement preconditioners do converge to a feasible point lying on the constraint manifold.

I will also describe a method which uses one application of a particular constraint preconditioner – which is independent of the Newton iteration – to project the approximate solution onto the constraint manifold while not destroying the progress that has been made by *any* iterative solver. Applying this technique removes all restrictions on the types of iterative methods which can be used safely in constrained optimization problems, and we hope will lead to the development of more high-quality preconditioners tailored to this important field.

I will illustrate the theoretical results with computations involving large-scale example problems taken from the CUTEst optimization testing environment.

Rational Orthogonal Functions and Rational Gauss Quadrature with Applications in Linear Algebra

Carl Jagels, Miroslav Pranić and Lothar Reichel

Abstract

It is well known that polynomials that are orthogonal with respect to a nonnegative measure on an interval on the real axis satisfy a three-term recurrence relation. This property of orthogonal polynomials is the basis for many popular and efficient algorithms in numerical linear algebra, including the symmetric Lanczos process for partial tridiagonalization of a large symmetric matrix and Golub–Kahan bidiagonalization for partial reduction of a large nonsymmetric matrix to bidiagonal form. These methods are important for the computation of a few eigenvalues and associated eigenvectors of a large symmetric matrix, for the computation of a few singular values and associated singular vectors of a large matrix, and for the approximation of matrix functions. Properties of the symmetric Lanczos process and Golub–Kahan bidiagonalization can be described in terms of suitable Krylov subspaces. These methods therefore are related to polynomial approximation.

In some contexts, such as when one would like to compute a few eigenvalues in a specified interval and associated eigenvectors of a large symmetric matrix, the application of a rational Lanczos method can yield the desired eigenpairs with less computational effort than the standard Lanczos method. The reason for this is that rational Lanczos methods use rational Krylov subspaces, which for a suitable choice of poles can yield much faster convergence than standard Krylov methods. Rational Krylov subspace methods also can be attractive to use for the approximation of matrix functions.

Similarly as orthogonal polynomials, orthogonal rational functions with specified poles can satisfy short recurrence relations. The recursion relations are more complicated for orthogonal rational functions than for orthogonal polynoials, because they depend on the number of distinct poles and the order in which they are used. The fact that orthogonal rational functions with poles at zero and infinity, and whose numerator and denominator degrees are increased alternatingly, satisfy short recursion relations has been known for some time; see Njåstad and Thron [5] as well as the review Jones and Njåstad [4]. This result was rediscovered by Simoncini fairly recently in a linear algebra context in a paper that contains many interesting ideas [7].

It is interesting to allow the numerator degree of the orthogonal rational functions with poles at zero and infinity to grow faster than the denominator degree, because in the rational Lanczos method each increase in the numerator degree requires the evaluation of a matrix-vector product with a large symmetric matrix A, while each increase of the denominator degree demands the solution of a linear system of equations with this matrix. The latter task can be much more time consuming than the former. The recursion formulas for this kind of rational Lanczos process have been studied in [2]. In particular, it is shown that short recursion formulas are valid also in this situation. The recursion formulas depend not only on the measure defined by the matrix A and the initial vector b, but also on how frequently the denominator degree is increased.

Also when the rational Lanczos process is determined by several (finite or infinite) real or complex conjugate poles, the recursion relations for the rational Lanczos process can be short. The number of terms in the recursion relations depends on the number of distinct poles and their ordering. Recursion relations for this kind of orthogonal rational functions have recently been investigated in [6]. These orthogonal rational functions are well suited for the approximation of matrix functions of the form $f(A)v$, where A is a large symmetric matrix and v is a vector. We will review properties

of the recursion relations for orthogonal rational functions and discuss the application of these functions to the approximation of matrix functions.

Orthogonal polynomials are related to Gauss quadrature rules, and orthogonal rational functions are related to rational Gauss quadrature rules in a similar way. The latter rules are exact for certain classes of rational functions with prescribed poles. The zeros of orthogonal rational functions are nodes of rational Gauss rules. They can be computed as the eigenvalues of the symmetric banded matrix determined by the recursion relations. The square of the first components of suitably scaled eigenvectors of this matrix give the weights of the quadrature rules. These relations appear to be new. It follows that the nodes and weights of rational Gauss rules can be computed by a Golub–Welsch-type algorithm. We will describe this algorithm.

Rational Gauss–Radau rules, which are rational Gauss rules with one preselected node, also can be computed by a Golub–Welsch-type algorithm. Under suitable conditions, pairs of rational Gauss and Gauss–Radau rules provide upper and lower bounds for matrix functions of the form $v^T f(A)v$, where v is a vector. These results generalize techniques for (standard) Gauss and Gauss–Radau rules, described, e.g., by Golub and Meurant [1], to rational analogues. Some results can be found in [3], more will be presented together with applications.

Orthogonal rational functions with respect to a measure in the complex plane also satisfy certain recursion relations that find application in numerical linear algebra. Time permitting, we will discuss their properties and some applications.

References

[1] G. H. Golub and G. Meurant, Matrices, Moments and Quadrature with Applications, Princeton University Press, Princeton, 2010.

[2] C. Jagels and L. Reichel, Recursion relations for the extended Krylov subspace method, Linear Algebra Appl., 434 (2011), pp. 1716–1732.

[3] C. Jagels and L. Reichel, The structure of matrices in rational Gauss quadrature, Math. Comp., 82 (2013), pp. 2035–2060.

[4] W. B. Jones and O. Njåstad, Orthogonal Laurent polynomials and strong moment theory: a survey, J. Comput. Appl. Math., 105 (1999), pp. 51–91.

[5] O. Njåstad and W. J. Thron, The theory of sequences of orthogonal L-polynomials, in Padé Approximants and Continued Fractions, eds. H. Waadeland and H. Wallin, Det Kongelige Norske Videnskabers Selskab, Skrifter 1, 1983, pp. 54–91.

[6] M. S. Pranić and L. Reichel, Recurrence relations for orthogonal rational functions, Numer. Math., 123 (2013), pp. 629–642.

[7] V. Simoncini, A new iterative method for solving large-scale Lyapunov matrix equations, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.

Resolution Arguments for the Estimation of Regularization Parameters in the Solution of Ill-Posed Problems

Jakob Hansen, Michael Horst and Rosemary Renaut

Abstract

The solution of the ill-posed linear system of equations, $A\mathbf{x} = \mathbf{b}$, where A is of size $m \times n$ is considered. Tikhonov regularization is a standard technique for finding an *acceptable* solution with given certain smoothness properties through seeking the solution of the system augmented with the constraint $\|L\mathbf{x}\|_2 \leq \delta$ for a chosen matrix L of size $p \times n$ and a noise dependent δ. We consider the general formulation

$$\mathbf{x} = \arg\min\{\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2\|L(\mathbf{x} - \mathbf{x_0})\|_2^2\}, \tag{1}$$

which includes the prior information $\mathbf{x_0}$ and the unknown regularization parameter λ. Our focus is on the estimation of the parameter λ when the system is obtained by downsampling of a given model where m and n are large. In particular we replace (1) by the sequence of problems

$$\mathbf{x_k} = \arg\min\{\|A_k\mathbf{x} - \mathbf{b_k}\|_2^2 + {\lambda_k}^2\|L_k(\mathbf{x} - (\mathbf{x_k})_0)\|_2^2\}, \tag{2}$$

where the subscript denotes the downsampled system(s), measurements and prior information. Parameter estimation techniques for λ have been extensively studied e.g. [4, 5] and include standard approaches such as the L-curve, Generalized Cross Validation (GCV), Unbiased Predictive Risk (UPRE) and more recently analysis of the residual or augmented residual through the use of the residual periodogram and χ^2 distribution, respectively. For all approaches, other than the χ^2 distribution, the algorithm relies on the solution of (1) for a possibly large range of values for λ, and it is therefore essential that the solution is found efficiently. The χ^2 discrepancy, based on the χ^2 distribution for the functional, on the other hand, leads immediately to a Newton root finding algorithm for λ which is typically solved in about 10 steps, dependent on the knowledge of noise in the measurements $\mathbf{b}$. For a large system of equations, for example with image restoration, or three dimensional image reconstruction, even the use of 10 solves may be too many. Here we will discuss an alternative direction through the solution of the downsampled problems (2). The analysis relies on demonstrating the convergence of $\lambda_k \xrightarrow{k} \lambda_n$ when λ_n is obtained through the χ^2 technique, and that information on the covariance of the noise in the measurements is available.

We present the results of our analysis for two cases. The first when A is obtained from the discretization of an integral equation with a square integrable kernel. In this case we appeal to results of Hansen [5, 3] on the relationship of the continuous singular value expansion to the singular value decomposition. Because the χ^2 discrepancy result can be applied using the SVD, when $L = I$, an explicit formulation arises that yields the necessary convergence. While one may have the impression that the SVD is of limited use for large scale problems, we note that recent advances on the development of randomized SVDs reassert the relevance of solutions that can take advantage of this framework [2, 9], noting also that the χ^2 result still applies with appropriate modification in the context of a truncated SVD [7]. The second case is specifically for kernels that permit implementation using a discrete Fourier or cosine transform (DFT, DCT) , as for example with image deblurring and periodic or reflexive boundary conditions, in which cases standard convergence results for the DFT are employed. Further, if L also admits an appropriate kernel, the approach goes beyond the case $L = I$, and hence has application for other standard smoothing norms.

In providing these results for the χ^2 discrepancy for finding λ in (2), we note that it is very specifically the existence of a unique estimate for λ_k which permits the downsampling approach. On the other hand, parameters obtained by techniques such as GCV and UPRE, which are better known for parameter estimation, are less well characterized numerically and do not appear to lend themselves to the downsampling. Still, within this work we present an analysis which shows that **a** minimum of the UPRE functional is provided by **the** parameter obtained via the χ^2 algorithm, and in this sense can also see the result as satisfying the principle of unbiased risk for the solution.

Supporting numerical results will be provided, demonstrating that one can achieve satisfactory solutions. This work contrasts with approaches that seek to find the solution of a large scale problem by projection to a smaller subspace, directly from the large scale formulation, as for example in the use of the iterative LSQR with generalized cross validation regularization on the subproblem, e.g. [1] or other hybrid techniques, [6, 8]. Discussion of the relevance of the analysis for the underdetermined case will also be provided.

References

[1] J. Chung, J. Nagy and D.P. O'Leary, (2008), *A weighted GCV method for Lanczos hybrid regularization*, ETNA, 28, 149-167.

[2] N. Halko, P. G. Martinsson, and J. A. Tropp, (2011), *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Review, 53(2), 217-288.

[3] P. C. Hansen, *Computation of the singular value expansion, Computing*, 40, 3, 185-199, 1988.

[4] P. C. Hansen, (1998), *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, (SIAM Monographs on Mathematical Modeling and Computation 4), Philadelphia.

[5] P. C. Hansen, *Discrete Inverse Problems, Insight and Algorithms*, SIAM, Philadelphia, 2010.

[6] I. Hnetynkova, M. Plesinger and Z. Strakos, (2009), *The regularizing effect of the Golub-Kahan iterative bidiagonalization and revealing the noise level in the data*, BIT, Numer. Math, **49** 669-696.

[7] J. L. Mead and R. A. Renaut (2009), *A Newton root-finding algorithm for estimating the regularization parameter for solving ill-conditioned least squares problems*, *Inverse Problems*, **25**, 025002, doi: 10.1088/0266-5611/25/2/025002.

[8] R. Renaut, I. Hnetynkova and J. Mead (2010), *Regularization parameter estimation for large scale Tikhonov regularization using a priori information*, *Computational Statistics and Data Analysis*, **54**, 12, 3430-3445, doi:10.1016/j.csda.2009.05.026.

[9] V. Rokhlin and M. Tygert, (2008), *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Natl. Acad. Sci. USA, **105**, 13212-13217.

Numerical Behavior of Indefinite Orthogonalization

Miro Rozložník, Felicja Okulicka-Dłużewska and Alicja Smoktunowicz

Abstract

For a real symmetric nonsingular matrix $B \in \mathcal{R}^{m,m}$ and for a full column rank matrix $A \in \mathcal{R}^{m,n}$ ($m \geq n$) we look for the decomposition $A = QR$, where the columns of $Q \in \mathcal{R}^{m,n}$ are mutually orthogonal with respect to the bilinear form induced by the matrix B so that $Q^T BQ = \Omega = \mathrm{diag}(\pm 1)$ and where $R \in \mathcal{R}^{n,n}$ is upper triangular with positive diagonal elements. Such problems appear explicitly or implicitly in many applications such as eigenvalue problems, matrix pencils and structure-preserving algorithms, interior-point methods or indefinite least squares problems. It is clear that for $B = I$ or $B = -I$ we get the standard QR decomposition of the matrix A. If B is positive and diagonal then the problem is equivalent to the standard decomposition of the row-scaled matrix $\mathrm{diag}^{1/2}(B)A$. For a general but still symmetric positive definite B, the factors can be obtained from the QR factorization in the form $B^{1/2}A = (B^{1/2}Q)R$. The indefinite case of $B \in \mathrm{diag}(\pm 1)$ has been studied extensively by several authors. It appears that under assumption on nonzero principal minors of $A^T BA$ each nonsingular square A can be decomposed into a product $A = QR$ with $Q^T BQ \in \mathrm{diag}(\pm 1)$ and R being upper triangular. These concepts can be extended also to the case of a full column rank A and a general indefinite (but nonsingular) matrix B.

Although all orthogonalization schemes are mathematically equivalent, their numerical behavior can be significantly different. The numerical behavior of orthogonalization techniques with the standard inner product $B = I$ has been studied extensively over last several decades including the Householder, Givens QR and modified Gram-Schmidt. The classical Gram-Schmidt (CGS) and its reorthogonalized version have been studied much later in [1, 4, 5]. It is also known that the weighted Gram-Schmidt with diagonal B is numerically similar to the standard process applied to the row-scaled matrix $\mathrm{diag}^{1/2}(B)A$. Several orthogonalization schemes with a non-standard inner product have been studied including the analysis of the effect of conditioning of B on the factorization error and on the loss of B-orthogonality between the computed vectors [2].

In this contribution we consider the case of symmetric indefinite B and assume that $A^T BA$ is strongly nonsingular (i.e. that each principal submatrix $A_j^T BA_j$ is nonsingular for $j = 1, \ldots, n$, where A_j denotes the matrix with the first j columns of A). Then the Cholesky-like decomposition of indefinite $A^T BA$ exists and the triangular factor R can be recovered from $A^T BA = R^T \Omega R$. We first analyze the conditioning of factors Q and R. It is clear that if B is positive definite then $\|R\| = \|B^{1/2}A\|$, $\|R^{-1}\| = 1/\sigma_{min}(B^{1/2}A)$ and $\|Q\| \leq \|B^{-1}\|^{1/2}$, $\sigma_{min}(Q) \geq 1/\|B\|^{1/2}$. Therefore $\kappa(R) = \kappa(B^{1/2}A) = \kappa^{1/2}(A^T BA)$ and $\kappa(Q) \leq \kappa^{1/2}(B)$. However, for B indefinite we have only $\|A^T BA\| \leq \|\Omega\|\|R\|^2$ and $\|(A^T BA)^{-1}\| \leq \|\Omega^{-1}\|\|R^{-1}\|^2$ and so the square root of the condition number of $A^T BA$ is just a lower bound for the condition number of the factor R, i.e. $\kappa^{1/2}(A^T BA) \leq \kappa(R)$. The upper bound for $\kappa(R)$ seems more difficult to obtain. One must look at its principal submatrices R_j and derive bounds for the norm of their inverses $\|R_j^{-1}\|$ considering

$$(R_j^T R_j)^{-1} = \begin{pmatrix} (R_{j-1}^T R_{j-1})^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \omega_j \left[(A_j^T BA_j)^{-1} - \begin{pmatrix} (A_{j-1}^T BA_{j-1})^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right].$$

This identity provides the basic insight into relation between the minimum singular value of the factor R and the minimum singular values of some principal submatrices of $A_j^T BA_j$. Observe that its recursive use leads to the expansion of the matrix $(R^T R)^{-1}$ in terms of $(A^T BA)^{-1}$ and in terms of only those inverses of principal submatrices $(A_j^T BA_j)^{-1}$ where there is a change of the sign in

the factor $\Omega = diag(\omega_1, \ldots, \omega_n)$, i.e. for such $j = 1, \ldots, n-1$ where $\omega_{j+1} \neq \omega_j$. It follows then that $|\omega_{j+1} - \omega_j| = 2$ and therefore we have the bound

$$\|R^{-1}\|^2 \leq \|(A^T BA)^{-1}\| + 2 \sum_{j;\ \omega_{j+1} \neq \omega_j} \|(A_j^T BA_j)^{-1}\|. \tag{1}$$

The norm of the factor R can be either bounded as $\|R\| \leq \|A^T BA\| \|R^{-T}\|$ or one can consider similar identity for $R^T R$ and get the bound for $\|R\|$ in terms of the Schur complements corresponding only to those principal submatrices $A_j^T BA_j$ (subject to $A^T BA$) where $\omega_{j+1} \neq \omega_j$, i.e.

$$\|R\|^2 \leq \|A^T BA\| + 2 \sum_{j;\ \omega_{j+1} \neq \omega_j} \|(A^T BA) \backslash (A_j^T BA_j)\|. \tag{2}$$

The bounds (1) and (2) can be reformulated also for quasi-definite matrices $A^T BA$, where the factor Ω has a particular structure $\Omega = diag(I; -I)$ with appropriate dimensions. The only nonzero term in the sum over principal submatrices corresponds to the biggest positive definite principal submatrix of $A^T BA$. The singular values of the factor Q can be bounded from $Q = AR^{-1}$.

Here we analyze two types of important schemes used for orthogonalization with respect to the bilinear form induced by B. We give the worst-case bounds for quantities computed in finite precision arithmetic and formulate our results on the loss of orthogonality and on the factorization error (measured by $\|\bar{Q}^T B\bar{Q} - I\|$ and $\|A - \bar{Q}\bar{R}\|$) in terms of quantities proportional to the roundoff unit u, in terms of $\|A\|$ and $\|B\|$ and in terms of the extremal singular values of factors $\bar{Q}$ and $\bar{R}$. Based on previous discussion the latter depend on the extremal singular values of the matrix $A^T BA$ and principal submatrices $A_j^T BA_j$ with the change of the sign $\bar{\omega}_{j+1} \neq \bar{\omega}_j$ during the orthogonalization. First we analyze the QR implementation based on the Cholesky-like decomposition of indefinite $A^T BA$. We show that assuming $\mathcal{O}(u)\kappa(A^T BA)\|A\|^2\|B\| \max_{j,\ \bar{\omega}_{j+1} \neq \bar{\omega}_j} \|(A_j^T BA_j)^{-1}\| < 1$ such decomposition runs to completion and the computed factors $\bar{R}$ and $\bar{\Omega}$ satisfy $A^T BA + \Delta B = \bar{R}^T \bar{\Omega} \bar{R}$ with $\|\Delta B\| \leq \mathcal{O}(u)[\|\bar{R}\|^2 + \|B\|\|A\|^2]$. For the computed orthogonal factor $\bar{Q}$ it follows then that $\|\bar{Q}^T B\bar{Q} - \bar{\Omega}\| \leq \mathcal{O}(u)\kappa(\bar{R}) \left[\kappa(\bar{R}) + 2\|B\bar{Q}\|\|\bar{Q}\|\right]$. The accuracy of these factors can be improved by one step of iterative refinement when we apply the same decomposition to the actual $\bar{Q}^T A\bar{Q}$ and get the bound $\|\bar{Q}^T B\bar{Q} - \bar{\Omega}\| \leq \mathcal{O}(u)\|B\|\|\bar{Q}\|^2$. We consider also the B-CGS algorithm and its version with reorthogonalization and show that their behavior is similar to Cholesky-like QR decomposition and its variant with refinement, respectively. The details can be found in [3].

References

[1] L. Giraud, J. Langou, M. Rozložník, and J. van den Eshof. Rounding error analysis of the classical Gram-Schmidt orthogonalization process, Num. Math., 101: 87-100, 2005.

[2] M. Rozložník, J. Kopal, M. Tuma, A. Smoktunowicz, Numerical stability of orthogonalization methods with a non-standard inner product, BIT Numerical Mathematics (2012) 52:1035-1058.

[3] M. Rozložník, A. Smoktunowicz and F. Okulicka-Dłużewska. Indefinite orthogonalization with rounding errors, 2013, to be submitted.

[4] A. Smoktunowicz, J.L. Barlow, and J. Langou. A note on the error analysis of classical Gram-Schmidt. Numer. Math., 105(2):299–313, 2006.

[5] J. Barlow and A. Smoktunowicz. Reorthogonalized block classical Gram–Schmidt, Numerische Mathematik 123 (3), 395–423 (2013).

The Two Sided Arnoldi Algorithm

Axel Ruhe

Abstract

At the Matrix Pencil Conference in Piteå1982, I described a two sided Arnoldi algorithm to compute approximative left and right eigenvectors, y' and x, of a large nonsymmetric matrix B, [2]. After that, I moved to Göteborg and forgot all about left eigenvectors, until I did send the paper [3] by Wu, Wei, Jia, Ling and Zhang to print for BIT, and noted that it contained a perturbation analysis for the same nonsymmetric eigenproblem in the spirit of the Kahan, Parlett and Jiang paper [1].

In the talk, I will report on some tests of this algorithm, [2],using tools that were not available 30 years ago. The aim is to compare the two sided Arnoldi algorithm to the well studied, but little used, nonsymmetric Lanczos algorithm, and the obvious choice of doing two runs of the widely implemented standard Arnoldi algorithm, first on B and then on B'.

References

[1] Kahan, W., Parlett, B.N., Jiang, E.: Residual bounds on approximate eigensystems of nonnormal matrices. SIAM Journal on Numerical Analysis **19**, 470–484 (1982)

[2] Ruhe, A.: The two-sided Arnoldi algorithm for nonsymmetric eigenvalue problems. In: B. Kågström, A. Ruhe (eds.) Matrix Pencils, LNM 973, pp. 104–120. Springer-Verlag, Berlin Heidelberg New York (1983)

[3] Wu, G., Wei, Y., Jia, Z., Ling, S., Zhang, L.: Towards backward perturbation bounds for approximate dual Krylov subspaces. BIT Numerical Mathematics **53**(1), 225–239 (2013)

Wilkinson-Type Error Bounds Revisited

Claude-Pierre Jeannerod and Siegfried M. Rump

Abstract

Wilkinson's standard error bounds for floating-point summation and dot products involve a factor $\gamma_k := k\mathbf{u}/(1-k\mathbf{u})$, where $\mathbf{u}$ denotes the relative rounding error unit. Here the denominator covers higher order terms.

Surprisingly this factor can be replaced by $k\mathbf{u}$ [4, 2]. It applies to standard floating-point arithmetic (in base β, precision p) and barring underflow and overflow. The estimates are valid no matter what the order of evaluation and without restriction on k.

Moreover, the famous Lemma 8.4 in Higham's ASNA [1] bounds the error of Gaussian elimination of an $n \times n$ matrix by factors γ_k for $k = n$ or $k = n-1$, depending on whether division occurs or not. These factors can be replaced by $k\mathbf{u}$ as well [3]. A similar remark applies to Cholesky factorization and solving triangular systems by substitution. Again, all estimates are true no matter what the order of evaluation and without restriction on k.

One might hope to replace γ_k generally by $k\mathbf{u}$ for floating-point error bounds of arithmetic expression trees of depth k. However, this is not true, as for example for binary summation - at least for huge exponent range and huge k.

References

[1] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, second ed., 2002.

[2] C.-P. Jeannerod and S. M. Rump, *Improved error bounds for inner products in floating-point arithmetic*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 338–344.

[3] S. M. Rump and C.-P. Jeannerod, *Improved backward error bounds for LU and Cholesky factorizations*, submitted for publication.

[4] S. M. Rump, *Error estimation of floating-point summation and dot product*, BIT, 52 (2012), pp. 201–220.

Using Partial Spectral Information for Block Diagonal Preconditioning of Saddle-Point Systems

Daniel Ruiz, Annick Sartenaer and Charlotte Tannier

Abstract

We consider the solution by an iterative solver of the (possibly large and sparse) saddle-point linear system

$$\mathcal{A}u = b \equiv \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, with $n \geq m$. We assume that A is symmetric and positive definite and that B has full column rank. Such kind of systems typically arise in constrained nonlinear optimization, as the result for first-order optimality conditions, where $\mathcal{A}$ is known as the Karush-Kuhn-Tucker (KKT) matrix. The assumption of positive-definiteness of A is met, in particular, when solving strictly convex quadratic optimization problems. On the application side, systems structured as (1) where A is naturally symmetric and positive definite arise in CFD or in magnetostatics for instance, from the numerical solution of PDEs or in PDE-constrained optimal control.

We also assume that A and B are possibly ill-conditioned, but that some form of first-level of preconditioning has been applied so that the spectrum of A as well as the singular value distribution of B are both condensed, with relatively few very small eigenvalues and singular values, respectively. Our goal is then to complete this first-level of preconditioning applied separately on A and B in an appropriate manner, so that efficient iterative solutions of the system (1) is ensured despite these trailing sets of extreme eigenvalues/singular values. To this aim, and following our approach, we consider that some good approximation of the spectral information associated to the smallest eigenvalues in A and to the smallest singular values in B is available. In this talk, we will focus on the case where only some good approximation of the spectral information associated to the smallest eigenvalues in A is used (this can be achieved with techniques like those proposed in [1], for instance).

It is well understood from previous studies that it not sufficient to be able to solve independently the subproblems associated to the $(1,1)$ block and to the constraints normal equations B^TB to build efficient preconditioners $\mathcal{P}$ for matrix $\mathcal{A}$ in (1), and that it is important to combine A with B in some way. We shall show how it is possible to combine the spectral information extracted from A to construct a cheap preconditioner $\mathcal{P}$ that ensures fast convergence in MINRES for the solution of system (1). The approach we propose is based on the study by Murphy, Golub and Wathen (see [2]), where the authors show that the "ideal" block diagonal preconditioner

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}, \tag{2}$$

in which the exact Schur complement, $S = B^TA^{-1}B$, is incorporated, yields a preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ with exactly three distinct eigenvalues (under the assumption of positive-definiteness for A). This property is of practical use when inexpensive approximations of the inverse of A and S are available. We will propose two variants for the block diagonal preconditioning matrix $\mathcal{P}$ in (2) based on a Schur complement approximation derived from the prior spectral information extracted from A directly, i.e., using some good approximation of the subspace associated to the

smallest eigenvalues in A. Based on the results provided by Rusten and Winter (see [3]), we study the spectral properties of the preconditioned matrix in both cases and give evidence how one can *recombine appropriately* the available spectral information from the (1,1) block, through the Schur complement approximation, to build an efficient block diagonal preconditioner with little extra cost. We then illustrate through some numerical experiments the benefits of this approach when solving system (1) with the preconditioned MINRES iterative method.

These considerations will lead us, in a second part of the talk, to highlight some aspects of the interaction between A and B when solving systems of the form (1). Indeed, it is commonly observed that despite their possible ill-conditioning, some sort of recombination of A and B occur that either spoil the convergence of Krylov subspace methods like MINRES, either not, but no clear insight on how and why this occurs exists, to our knowledge. We will give some insight, through the study of the Schur complement approximation based on the few very small eigenvalues, on how and in which circumstances the bad conditioning contained in these eigenvalues effectively spoils the convergence of MINRES.

References

[1] G. Golub, D. Ruiz, and A. Touhami: *A hybrid approach combining Chebyshev filter and conjugate gradient for solving linear systems with multiple right-hand sides.* SIAM Journal on Matrix Analysis and Applications, 2007, 29(3): pp. 774–795.

[2] M. Murphy, G. Golub, and A. Wathen: *A note on preconditioning for indefinite linear systems.* SIAM Journal on Scientific Computing, 2000, 21(6): pp. 1969–1972.

[3] T. Rusten and R. Winther: *A preconditioned iterative method for saddle-point problems.* SIAM Journal on Matrix Analysis and Applications, 1992, 13(3): pp. 887–904.

Computing the Rank and Nullspace of Rectangular Sparse Matrices

Nick Henderson, Ding Ma, Michael Saunders and Yuekai Sun

Abstract

To model biochemical networks, systems biologists are generating increasingly large *stoichiometric matrices* S, whose rows and columns correspond to chemical species and chemical reactions. An important step toward evaluating drug targets and analyzing transient behavior of the network is called *conservation analysis*, which reduces to finding the *rank of* S and the *nullspace of* S^T. SVD is not practical, but with care, sparse QR or LU factors can serve both purposes.

In general, rank-revealing factorizations of a matrix A must ensure that all factors except one are well-conditioned. The rank of the remaining factor then reflects the rank of the original matrix. SVD is the ideal example: $A = UDV^T$ with U and V orthogonal (perfectly conditioned). The rank of D accurately reflects the rank of A.

The SuiteSparseQR package of Davis [1, 2] with its MATLAB interface SPQR provides multithreaded multifrontal rank-revealing QR factors of a rectangular sparse matrix. For some large stoichiometric examples A, SPQR has proved remarkably efficient at computing factors $AP = QR$, where P is a sparsity-preserving column ordering chosen in advance, Q is stored as a product of sparse Householder matrices, and R is a sparse trapezoidal matrix. With P and Q perfectly conditioned, the rank of R must reflect the rank of A. The product form of Q provides an ideal basis for the nullspace of A^T. However, the storage requirements are sometimes rather high for QR factors, especially if A contains some rather dense rows. We therefore consider sparse LU factors.

LUSOL [3, 4] and its MATLAB interface [5] provide LU factors of a rectangular sparse matrix. The normal *threshold partial pivoting* option finds row and column permutations P_1, P_2 such that $P_1AP_2 = LU$, where L has unit diagonals and bounded off-diagonals and is therefore likely to be well-conditioned. The rank of U is not immediately obvious, but it reflects the rank of A.

LUSOL's *threshold rook pivoting* factorization is best thought of as $P_1AP_2 = LDU$, where both L and U have unit diagonals and bounded off-diagonals and are likely to be well-conditioned. The rank of D then reflects the rank of A. This option is reliable and the factors can be significantly more sparse than those from SPQR. Unfortunately, the LDU factors can be significantly more expensive to compute. We therefore study two alternative ways to use LUSOL. Starting with threshold *partial* pivoting factors $P_1AP_2 = LU$ (with L well-conditioned),

either: apply threshold *rook* pivoting to U,
or: apply threshold *partial* pivoting to U^T.

[1] T. A. Davis, SuiteSparseQR: multithreaded multifrontal sparse QR factorization, `http://www.cise.ufl.edu/research/sparse/SPQR/`.

[2] T. A. Davis, SuiteSparseQR: Algorithm 9xx: SuiteSparseQR, a multifrontal multithreaded sparse QR factorization package, submitted to *ACM TOMS*.

[3] P. E. Gill, W. Murray, and M. A. Saunders, SNOPT: An SQP algorithm for large-scale constrained optimization, *SIAM Rev.* 47(1), 99–131 (2005).

[4] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, Maintaining LU factors of a general sparse matrix, *Linear Alg. Appl.* 88/89, 239–270 (1987).

[5] N. W. Henderson, MATLAB interface to LUSOL, `https://github.com/nwh/lusol`.

Computing the Nearest Pencil $A - \lambda A^T$ without Unimodular Eigenvalues

Federico Poloni and Christian Schröder

Abstract

Real palindromic matrix pencils, i.e., pencils of the form $A - \lambda A^T$, arise in applications such as parameter estimation of econometric time series [2], or passive linear dynamical systems [3]. In both cases, the pencil is not allowed to have unimodular eigenvalues (i.e., ones on the unit circle) in order to be meaningful. In this case we will call the matrix A *admissible*. For example, in the mentioned applications the presence of unimodular eigenvalues indicates an unsolvable subproblem or the loss of passivity.

However, due to inexactness in the derivation of the pencil like linearization errors, modeling errors, omitting higher order terms, etc., it is well possible to end up with a (slightly) in-admissible A when it really should be admissible due to outer conditions (e.g., the physical system to be modeled is known to be passive, i.e., it cannot generate energy). In these cases, a regularization procedure is necessary that computes a perturbation E such that $A + E$ becomes admissible. Often this perturbation is restricted to lie in some linear subspace, i.e., to be of the form

$$E = \sum_{i=1}^{m} \delta_i E_i$$

for some given matrices $E_i, i = 1, 2, \ldots, m$ (e.g., E could be required to have a certain nonzero pattern [3]). The task is to determine the vector $\delta = [\delta_1, \ldots, \delta_m]^T$ of smallest 2-norm such that $A + E$ is admissible.

Note that eigenvalues of real palindromic pencils must come in quadruples $(\lambda, \overline{\lambda}, \lambda^{-1}, \overline{\lambda}^{-1})$. Here, $\overline{\lambda}$ denotes the complex conjugate of λ. These quadruples reduce to pairs on the real axis and on the unit circle. As a consequence, a simple unimodular eigenvalue cannot leave the unit circle under small perturbations unless it merges with another unimodular eigenvalue (much like a simple real eigenvalue of a real matrix cannot leave the real axis under small real perturbations unless it merges with another real eigenvalue).

A basic algorithm that makes the unimodular eigenvalues merge to then split off the unit circle was introduced in [4] and developed further in [2, 3]. It is based on first order eigenvalue perturbation theory.

The method to be presented in my talk is an enhanced version that

1. avoids dealing directly with close-to-each-other eigenvalues, treating instead their average, which is known to behave much less sensitive to small perturbations [1]
2. additionally aims to prevent the non-unimodular eigenvalues from entering the unit circle
3. looks for the smallest admissible perturbation of A, instead of just some small perturbation.

This results is an algorithm that, compared to the state of the art, is more stable, needs less iterations, and yields a smaller perturbation.

It should also be mentioned that it is straight forward to adapt the algorithm to quadratic palindromic eigenvalue problems, or to remove imaginary eigenvalues of a symmetric/skew-symmetric pencil – which are what [2, 3] are actually about.

References

[1] Z. Bai, J. Demmel, and A. McKenney, *On computing condition numbers for the nonsymmetric eigenproblem*, ACM Trans. Math. Software, vol. 19(2), pp. 202 – 223, 1993

[2] T. Brüll, F. Poloni, G. Sbrana, and C. Schröder, *Enforcing solvability of a nonlinear matrix equation and estimation of multivariate ARMA time series*, preprint 1027, The MATHEON research center, Berlin, Germany, 2013

[3] T. Brüll and C. Schröder, *Dissipativity Enforcement via Perturbation of Para-Hermitian Pencils*, IEEE Transactions on Circuits and Systems I, vol. 60(1), pp. 164 – 177, 2013

[4] S. Grivet-Talocia, *Passivity enforcement via perturbation of Hamiltonian matrices*, IEEE Transactions on Circuits and Systems I, vol. 51(9), 2004

Memory-Efficient Incomplete Factorizations for Sparse Symmetric Systems

Jennifer Scott and Miroslav Tuma

Abstract

An important class of preconditioners for large sparse symmetric linear systems of equations is represented by incomplete Cholesky (IC) factorizations, that is, factorizations of the form LL^T in which some of the fill entries that would occur in a complete factorization are ignored. More generally, incomplete LDL^T factorizations, with D block diagonal, are used. Over the last fifty years, many different algorithms for computing incomplete factorizations have been proposed and used to solve problems from a wide range of application areas.

Our aim is to design and develop a new robust and efficient general-purpose incomplete factorization package that uses a limited memory approach. The initial focus is on the positive-definite case. Our IC software package, `HSL_MI28` [3, 2], exploits ideas from the positive semidefinite Tismenetsky-Kaporin modification scheme [1, 5] and is based on a matrix decomposition of the form

$$A = (L + R)(L + R)^T - E,$$

where L is a lower triangular matrix with positive diagonal entries that is used for preconditioning, R is a strictly lower triangular matrix with small entries that is used to stabilize the factorization process, and the error matrix E has the structure

$$E = RR^T.$$

Through the incorporation of the intermediate matrix R, `HSL_MI28` offers a generalisation of the widely-used and well-known ICFS algorithm of Lin and Moré [2]. We allow both the sparsity density of the incomplete factor and the amount of memory used in its computation to be controlled by the user. Factorization breakdown is circumvented by the use of a global shift strategy. Using extensive numerical experiments involving a large set of test problems arising from a wide range of real-world applications, we demonstrate the significant advantage of employing a modest amount of intermediate memory and show that, with limited memory, high quality yet sparse general-purpose preconditioners are obtained.

We are also interested in the more challenging case of indefinite linear systems and, in particular, we want to use incomplete factorizations to assist in efficiently solving saddle-point systems $Kx = b$, where K is of the form

$$K = \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix}.$$

Here A is $n \times n$ symmetric positive definite, B is rectangular and of full rank, and C is $m \times m$ ($m \leq n$) symmetric positive semi-definite. We propose extending our limited memory approach in a number different of ways. We first show that, with minor modifications, `HSL_MI28` can be used to compute a signed incomplete Cholesky factorization of the form LDL^T, where the D has entries ± 1. Again, we incorporate a global shift strategy to avoid breakdown. Numerical results are presented for practical problems for which $C = 0$ or $C = -\delta I$, with δ small (the latter arise from interior-point methods). Again, we examine the effects of employing intermediate memory in the computation of the incomplete factorization.

To try and improve performance (by reducing the amount of fill in L and/or the number of iterations required to achieve the requested accuracy), we next consider the incorporation of numerical pivoting into the incomplete factorization. Again, using a limited memory approach, we modify our software to compute an incomplete factorization LDL^T, where D has diagonal blocks of order 1 and 2. The use of global shifts is obviated by the pivoting; the penalty is a more complicated code. Comparisons are made with our signed IC approach and also with the recent SYM-ILDL code of Chen and Liu (`http://www.cs.ubc.ca/~inutard/`). The effects of orderings and scalings on the preconditioner quality are investigated.

This work was partially supported by the projects P201/13-06684S and 108/11/0853 of the Grant Agency of the Czech Republic and by EPSRC grant EP/I013067/1.

References

[1] I. E. Kaporin. High quality preconditioning of a general symmetric positive definite matrix based on its $U^TU + U^TR + R^TU$ decomposition. *Numerical Linear Algebra with Applications*, 5:483–509, 1998.

[2] C.-J. Lin and J. J. Moré. Incomplete Cholesky factorizations with limited memory. *SIAM J. on Scientific Computing*, 21(1):24–45, 1999.

[3] J. A. Scott and M. Tuma. `HSL_MI28`: an efficient and robust limited memory incomplete Cholesky factorization code. Technical Report RAL-TR-2013-P-004, Rutherford Appleton Laboratory, 2013.

[2] J. A. Scott and M. Tuma. On positive semidefinite modification schemes for incomplete Cholesky factorization. Technical Report RAL-TR-2013-P-005, Rutherford Appleton Laboratory, 2013.

[5] M. Tismenetsky. A new preconditioning technique for solving large sparse linear systems. *Linear Algebra and its Applications*, 154–156:331–353, 1991.

The Finite Section Method for Computing Exponentials of Doubly-Infinite Skew-Hermitian Matrices

Meiyue Shao

Abstract

In a number of scientific applications, especially in quantum mechanics, it is desirable to compute $\exp(\mathrm{i}A)$ where A is a self-adjoint operator. In practice, the operator A is often given in discretized form, i.e., a doubly-infinite Hermitian matrix under a certain basis, and a finite diagonal block of $\exp(\mathrm{i}A)$ is of interest. Suppose the $(-m:m,-m:m)$ block of $\exp(\mathrm{i}A)$ is desired. A simple way to solve this problem is to compute the exponential of the $(-w:w,-w:w)$ block of A, where w is chosen somewhat larger than m, and then use its central $(2m+1)\times(2m+1)$ block to approximate the desired solution. In reference to similar methods for solving linear systems, we call this approach *finite section method*. The diagonal blocks $(-m:m,-m:m)$ and $(-w:w,-w:w)$ are called the *desired window* and the *computational window*, respectively.

Despite the simplicity of the finite section method, it is crucial to understand how large the computational window needs to be, and whether this truncation produces sufficiently accurate approximation to the true solution. These questions are relatively easy to answer for bounded banded matrices, where standard polynomial approximation technique can be applied. But it turns out that the finite section method can also be applied to certain unbounded banded matrices, and still produces reliable solutions. In this work we explain this phenomenon and establish the finite section method with error estimates for several classes of doubly-infinite Hermitian matrices.

In the first part of this work, we discuss the case when A is bounded and banded. By the Benzi-Golub theorem [1], the off-diagonal entries of $\exp(\mathrm{i}A)$ decay exponentially with respect to the distance to the main diagonal, i.e., there exist constants $K>0$ and $\rho\in(0,1)$ such that

$$\left|\left[\exp(\mathrm{i}A)\right]_{ij}\right| \le K\rho^{|i-j|}, \qquad (\forall i,\, j).$$

Making use of this decay property, as well as the identity

$$\exp[t(X+\Delta X)]-\exp(tX)=\int_0^t \exp[(1-s)X]\Delta X\exp[s(X+\Delta X)]\,ds,$$

we show that the error in the finite section method are localized around the corners of the computational window. Therefore, by choosing a suitably large computational window based on our error estimate, the solution can be computed with guaranteed accuracy. Moreover, the distance $w-m$ stays constant when m increases.

In the second part of this work, we analyze the decay property of $\exp(\mathrm{i}A)$ and the error of the finite section method for several classes of unbounded matrices. Since A is unbounded, the decay bound of $\left|\left[\exp(\mathrm{i}A_{(-w:w,-w:w)})\right]_{ij}\right|$ provided by the Benzi-Golub theorem always deteriorate as w increases. To overcome this difficulty, we use a technique from [2] to identify localized eigenvectors of $A_{(-w:w,-w:w)}$ and then establish an error estimate for the finite section method. As a byproduct of the finite section method, we show that under certain assumptions, the doubly-infinite matrix $\exp(\mathrm{i}A)$ has the same decay property as the finite matrices $\exp(\mathrm{i}A_{(-w:w,-w:w)})$. We also propose an adaptive strategy for estimating the size of the computational window. This strategy works well even when a priori error estimates are too pessimistic or not easy to compute.

Finally, we present several numerical experiments to demonstrate the effectiveness of the finite section method.

A preprint of this work is available as [3].

References

[1] M. Benzi and G. H. Golub. Bounds for the entries of matrix functions with applications to preconditioning. *BIT*, 39(3):417–438, 1999.

[2] Y. Nakatsukasa. Eigenvalue perturbation bounds for Hermitian block tridiagonal matrices. *Appl. Numer. Math.*, 62:67–78, 2012.

[3] M. Shao. On the finite section method for computing exponentials of doubly-infinite skew-Hermitian matrices. Technical Report Nr. 26.2013, MATHICSE, EPF Lausanne, 2013. Available from `http://mathicse.epfl.ch/page-68906-en.html`.

Randomized Methods for Computing Null Spaces, with Applications to Rank-deficient Linear Systems

Josef Sifuentes, Zydrunas Gimbutas and Leslie Greengard

Abstract

A variety of problems in numerical linear algebra involve rank-deficient matrices. One example is computing the null space, denoted $\mathcal{N}(A)$, of a matrix $A \in \mathbb{C}^{n\times n}$ or the eigenspace of a given eigenvalue λ, given by

$$\mathcal{N}(A - \lambda I).$$

Another example is the solution of rank-deficient linear systems

$$Ax = b,$$

where A is rank k-deficient but b is in the range of A. A third category consists of problems like that above, but for which a set of k additional constraints are known of the form: $C^*x = f$, where $C \in \mathbb{C}^{n\times k}$ and $[A^*\ C]^*$ is full-rank, and $f \in \mathbb{C}^k$.

In this talk, we describe a very simple, deterministic framework for solving such problems, using *randomized* schemes. It is also worth noting that, in recent years, the use of randomization together with numerical rank-based ideas has proven to be a powerful combination for a variety of problems in linear algebra (see, for example, [1, 2, 3]).

The basic idea is remarkably simple and illustrated by the following example. Suppose we are given a rank-1 deficient matrix A and that we carry out the following procedure:

1. Choose a random vector $x \in \mathbb{C}^n$ and compute $b = Ax$.
2. Choose random vectors $p, q \in \mathbb{C}^n$ and solve
$$(A + pq^*)y = b$$
3. Then the difference $x - y$ is in the nullspace of A with probability 1.

In order for $A+pq^*$ to be invertible, we must have that $p \notin \mathcal{R}(A)$ and $q \notin \mathcal{R}(A^*)$. Since p and q are random, this must occur with probability 1. Consider the affine space, $\mathcal{S}_b = x' + \mathcal{N}(A)$ consisting of solutions to $Az = b$, where, $x' \in \mathcal{R}(A^*)$ is the solution of minimal norm. Then the difference of any two vectors in $\mathcal{S}_b$ lies in the nullspace of A. If $A + pq^*$ is nonsingular, y is the unique vector in $\mathcal{S}_b$ orthogonal to q, implying that $x - y \in \mathcal{N}(A)$.

This talk gives a rigorous framework for using this idea to compute null spaces as well as solve rank-deficient linear systems with or without additional constraints that would make the solution unique. Furthermore we will demonstrate that if A is well conditioned on a particular subspace of $\mathcal{C}^{n\times n}$, then the perturbed nonsingular matrix $A + PQ'$ is also well conditioned, allowing for the use of Krylov subspace based iterative methods to solve $(A + PQ')y = b$ when A is sparse or allows for fast matrix vector multiplications.

References

[1] N. Halko, P.G. Martinsson, and J. Tropp (2011), "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions", *SIAM Review*, **53**, 217–288.

[2] E. Liberty, F. Woolfe, P.G. Martinsson and M. Tygert (2007), "Randomized algorithms for the low-rank approximation of matrices", *PNAS*, **104**, 20167–20172.

[3] V. Rokhlin and M. Tygert (2008), "A fast randomized algorithm for overdetermined linear least-squares regression", *PNAS*, **105**, 13212–13217.

Fast Algorithms for Symmetric Tensor Contractions

Edgar Solomonik, Devin Matthews and James Demmel

Abstract

We give an algorithm that contracts symmetric tensors with a total of ω different indices in $\omega!$ fewer multiplications than a naive nonsymmetric algorithm. The cost of our algorithm is significantly lower than previously known methods, and yields some slightly faster symmetric matrix algorithms as arbitrarily faster contraction algorithms of general partially symmetric tensors.

We say that a tensor $\mathcal{A}$ is symmetric or "fully symmetric" if its value remains the same under any interchange of its indices $A_{i\ldots j\ldots} = A_{j\ldots i\ldots}$. We distinguish partially symmetric tensors, whose value remains the same under the interchange of only a subset or subsets of its indices. Pictorially, a symmetric tensor of dimension d and edges (rows, columns, etc.) of length n may be represented as a weighted hypergraph with n vertices and $\binom{n}{d}$ hyperedges, each containing d vertices.

Given any n-tuple $p = (i_1 \ldots i_n)$ we let $\chi_k(p)$ be the set of all pairs of a k-tuple and a $(n-k)$-tuple which are an ordered partition of p (e.g. $((i_1, i_3 \ldots), (i_2, i_4 \ldots)) \in \chi_k(p)$). A contraction between symmetric tensors $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, $\mathcal{C} = \mathcal{A} \odot \mathcal{B}$ may be written in the form

$$C_{i_1\ldots i_{s+t}} = \sum_{((j_1\ldots j_s),(l_1\ldots l_t))\in\chi_s(i_1\ldots i_{s+t})} \left(\sum_{k_1\ldots k_v} A^{k_1\ldots k_v}_{j_1\ldots j_s} \cdot B^{l_1\ldots l_t}_{k_1\ldots k_v} \right),$$

for some $s, t, v \geq 0$, we also denote $\omega = s + t + v$. We note that the symmetric contraction operator $\odot$ is commutative. Typically, such symmetric contractions are performed by (implicitly) forming the partially symmetric intermediate

$$\bar{C}^{l_1\ldots l_t}_{j_1\ldots j_s} = \sum_{k_1\ldots k_v} A^{k_1\ldots k_v}_{j_1\ldots j_s} \cdot B^{l_1\ldots l_t}_{k_1\ldots k_v}.$$

$\mathcal{C}$ may subsequently be obtained from $\bar{\mathcal{C}}$ by symmetrization of the l and j index groups. However, forming the $\bar{\mathcal{C}}$ intermediate requires $\frac{n^\omega}{s!t!v!}$ multiplications, since symmetry is preserved within each of three groups of indices in this representation.

Our main result is an algorithm that performs any such contraction using only $\binom{n}{\omega}$ multiplications to leading order. The new algorithm is based on the idea of forming a fully symmetric intermediate quantity and typically requires slightly more additions. It assumes the element-wise multiplications are commutative. The algorithm forms three fully symmetric intermediate quantities, the latter two of which can be computed with low-order cost,

$$Z_{i_1\ldots i_\omega} = \left(\sum_{((j_1\ldots j_s),(k_1\ldots k_v))\in\chi_s(i_1\ldots i_\omega)} A^{k_1\ldots k_v}_{j_1\ldots j_s} \right) \cdot \left(\sum_{((l_1\ldots l_t),(k_1\ldots k_v))\in\chi_t(i_1\ldots i_\omega)} B^{l_1\ldots l_t}_{k_1\ldots k_v} \right)$$

$$W_{i_1\ldots i_{\omega-1}} = \left(\sum_{((j_1\ldots j_s),(k_1\ldots k_v))\in\chi_s(i_1\ldots i_{\omega-1})} A^{k_1\ldots k_v}_{j_1\ldots j_s} \right) \cdot \left(\sum_{((l_1\ldots l_t),(k_1\ldots k_v))\in\chi_t(i_1\ldots i_{\omega-1})} B^{l_1\ldots l_t}_{k_1\ldots k_v} \right)$$

$$V_{i_1\ldots i_{\omega-1}} = \left(\sum_{((j_1\ldots j_s),(k_1\ldots k_{v-1}))\in\chi_s(i_1\ldots i_{\omega-1})} \sum_{k_v} A^{k_1\ldots k_v}_{j_1\ldots j_s} \right) \cdot \left(\sum_{((l_1\ldots l_t),(k_1\ldots k_{v-1}))\in\chi_t(i_1\ldots i_{\omega-1})} \sum_{k_v} B^{l_1\ldots l_t}_{k_1\ldots k_v} \right)$$

$$C_{i_1\ldots i_{s+t}} = \sum_{k_1\ldots k_v} Z_{i_1\ldots i_{s+t},k_1\ldots k_v} - n \cdot \sum_{k_1\ldots k_{v-1}} W_{i_1\ldots i_{s+t},k_1\ldots k_{v-1}} - \sum_{k_1\ldots k_{v-1}} V_{i_1\ldots,i_{s+t},k_1\ldots k_{v-1}}.$$

We first note some special cases of our result. For the multiplication of a symmetric matrix by a vector or a matrix (Basic Linear Algebra Subroutines (BLAS) `symv` and `symm`) we can use the symmetric contraction algorithm with $s = 1$, $t = 0$, and $v = 1$ to yield an algorithm that requires half the number of multiplications with respect to the standard method used in BLAS. For the symmetric rank-2k outer-product $A \cdot B^T + B \cdot A^T$, which corresponds to BLAS routines `syr2` and `syr2k`, our symmetric contraction algorithm (with $s = 1$, $t = 1$, and $v = 0$) also reduces the number of necessary multiplications by a factor of two. For the routines `symm` and `syr2k` our algorithm requires as many extra scalar additions as the number of scalar multiplications reduced. Therefore, when multiplications are more expensive than adds, the new algorithms are faster, which is the case in for instance, complex arithmetic, where each multiplication requires 6 operations and addition 2 operations.[1] For Hermitian matrix LAPACK routines `hemm`, `her2k`, and `heevd`, our symmetric contraction algorithm reduces the computational cost by about 1.4X over current methods. We note, however, that the algorithm has different numerical properties due to forming a different intermediate quantity. The numerical error bounds are asymptotically the same, but somewhat higher floating point error is accumulated for certain inputs.

Our construction also yields faster contraction algorithms for arbitrary partially-symmetric tensors. The extension is direct from nested use of the mathematical definition of a symmetric tensor contraction, which is possible because symmetric contractions are a commutative operator and each element-wise operation is assumed to be commutative by our algorithm. Note that when $\omega = 1$, the symmetric tensor contraction reduces to a nonsymmetric 1-dimensional contraction over an index which is shared by two tensors. Therefore, we can express any contraction of nonsymmetric tensors as nested symmetric contractions over a single index (with $\omega = 1$). Further, any tensor contraction where each tensor has partial-symmetry can be defined as a series of nested fully symmetric contractions. Since the symmetric tensor contractions each level is commutative, our fast symmetric algorithm can be used at each level in a nested fashion, yielding a potential benefit if any permutational index symmetry exists in the contraction.

As an example, consider the following contraction of partially symmetric tensors $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ (with all edge lengths still set to n),

$$C_{ij}^{mn} = \sum_{kl} A_{ij}^{kl} \cdot B_{kl}^{mn} + A_{mj}^{kl} \cdot B_{kl}^{in},$$

where each tensor has a single partial symmetry: $A_{ij}^{kl} = A_{kj}^{il}$, $B_{kl}^{mn} = B_{ml}^{kn}$, $C_{ij}^{mn} = C_{mj}^{in}$. We can write this as a nested symmetric tensor contraction $\mathbf{C}_{ij} = \sum_k \mathbf{A}_{ik} \otimes \mathbf{B}_{kj}$, where the operator $\otimes$ is defined to be three nested symmetric 1-dimensional contractions for each nonsymmetric index j, l, and n. These three nested symmetric contractions are a (commutative) representation of nonsymmetric matrix multiplication which gives the operator $\otimes$ a cost of $O(n^3)$, while matrix addition $(+)$ costs $O(n^2)$. Therefore, the reduction in the number of multiplications in the top level symmetric contraction over the lower indices would reduce the total number of operations for this tensor contraction by $\frac{(s+t+v)!}{s!t!v!} = 6$ (since $s = t = v = 1$ for the contraction of fully symmetric matrices).

Anti-symmetric (skew-symmetric) contractions have similar properties and all techniques presented here are extensible to this important case. In particular, tensors in highly accurate electronic structure calculation methods such as Coupled Cluster (used in the field of quantum chemistry) make extensive use of tensor contractions amongst partially anti-symmetric tensors.

[1] We note that Gauss's trick can be used to reduce the number of scalar multiplications from 4 to 3 for the complex multiplication, but does not reduce the number of total operations due to requiring more addition operations.

Minimum Residual Methods for Shifted Linear Systems with General Preconditioning

Kirk M. Soodhalter

Abstract

We present here an technique for solving a family (or a sequence of families) of linear systems in which the coefficient matrices differ only by a scalar multiple of the identity (shifted systems). The goal is to develop minimum residual methods for shifted systems compatible with general preconditioning

We parameterize the family by index i, i.e., the systems are of the form

$$Ax = b \quad \text{and} \quad (A + \sigma_i I)x(\sigma_i) = b, \quad i = 1, 2, \ldots, L, \tag{1}$$

with $A \in \mathbb{C}^{n\times n}$ and $\{\sigma_1, \ldots, \sigma_L\} \subset \mathbb{C}$. Such problems arise in many applications such as lattice quantum chromodynamics (QCD), Tikhonov-Phillips regularization, and Newton trust region methods. We can add an additional parameter j, indexing a sequence of matrices $\{A_j\} \subset \mathbb{R}^{n\times n}$ and for each j we solve a family of systems

$$A_j x = b_j \quad \text{and} \quad (A_j + \sigma_{i,j} I)x(\sigma_{i,j}) = b_j, \quad i = 1, 2, \ldots, L_j, \tag{2}$$

where $\{\sigma_{i,j}\}_{j=1}^{L_j} \subset \mathbb{C}$.

For sequences of "nearby" linear systems, subspace recycling techniques have been proposed, allowing important spectral information generated while solving $A_j x = b_j$ to be "recycled" and used to accelerate the convergence of the Krylov subspace iterative method used to solve $A_{j+1}x = b_{j+1}$; see, e.g., [Parks et al, SISC'05] and [Wang et al, IJNME'07].

Many methods have been proposed for solving (1), and they mostly are built upon the invariance of Krylov subspace under a shift of the coefficient matrix by a multiple of the identity, i.e.,

$$\mathcal{K}_j(A, r_0) = \mathcal{K}_j(A + \sigma I, r_0(\sigma)), \tag{3}$$

as long as the collinearity condition

$$r_0(\sigma) = \beta_0(\sigma) r_0 \text{ for some } \beta_0(\sigma) \in \mathbb{C}. \tag{4}$$

This relationship allows us to build Krylov subspace methods which simultaneously generate approximate solution corrections for all linear systems in (1) from the common Krylov subspace. These methods can be quite effective and allow for great savings in both storage and computational costs.

In [Soodhalter et al, Submitted 2013], we explored incorporating techniques for solving (1) into the framework of subspace recycling (for nonsymmetric matrices), specifically that of recycled GMRES (rGMRES) described in [Parks et al, 2005], in order to simultaneously solve systems in the family (2) for each parameter j while accelerating convergence by recycling spectral information. In the framework of recycled GMRES, we showed that it is in general not possible to maintain collinearity of the residuals for each system, a requirement which must hold each time we restart.

This highlights a downside of building methods for shifted systems upon the shift invariance property (3); we must maintain the residual collinearity. Furthermore, a more general difficulty is that methods built upon (3) are not compatible with arbitrary preconditioning, since in general

$$\mathcal{K}_j(M_1^{-1} A M_2^{-1}, M_1^{-1} r_0) \neq \mathcal{K}_j(M_1^{-1}(A + \sigma I)M_2^{-1}, M_1^{-1} r_0(\sigma)). \tag{5}$$

Therefore, we explore the possibility of solving either (1) or (2) by exploiting the shifted system structure without using the shift invariance (3). In the case of (2), we also seek a method which is compatible with the subspace recycling framework. By eliminating reliance on (3), we will be able to use general preconditioning.

Consider a nested sequence of search subspaces, $\mathcal{S}_1 \subset \cdots \subset \mathcal{S}_m \subset \cdots$ generated by a minimum residual iterative method (e.g., GMRES) applied to the unshifted system. For each shifted system, we apply the minimum residual Petrov-Gallerkin condition to compute all corrections from the same search subspace, i.e., at iteration m,

$$x_m(\sigma) = x_0(\sigma) + t_m(\sigma)$$

where $t_m(\sigma) \in \mathcal{S}_j$ but each residual is orthogonal to a different constraint space, i.e.,

$$r_m(\sigma) = b - (A + \sigma I)x_m(\sigma) \perp (A + \sigma I)\mathcal{S}_j. \tag{6}$$

For appropriate choices of $\mathcal{S}_j$, this can be accomplished inexpensively by exploiting the relationship between coefficient matrices of the shifted systems. In the case of solving a single family of shifted systems (1), $\mathcal{S}_m$ is simply the mth Krylov subspace, and we apply cycles of restarted GMRES to the unshifted system and inexpensively project the residuals for the shifted systems. In the case that we are solving a sequence of families of shifted systems (2), for each j, $\mathcal{S}_m$ is the mth augmented Krylov subspace and we are applying rGMRES to the unshifted system and again inexpensively projecting the shifted systems.

While producing approximations of improved quality for the shifted systems, this projection technique may not lead to convergence for the shifted system iterations. Upon convergence of the iteration for the unshifted system, this method can be called recursively for any remaining unconverged systems. This procedure does not require any collinearity relationship between the residuals. Thus it is completely compatible with general preconditioning. Incorporating preconditioning into this method is straightforward, though it does require some additional applications of the preconditioner.

Two methods can be derived from this strategy, one based on GMRES for solving (1) and one based on rGMRES for solving (2). We show how one can exploit the relationship of the shifted systems to perform these projections of the shifted system residuals inexpensively. Numerical results are presented demonstrating the effectiveness of this technique in both the unpreconditioned and preconditioned cases.

The strategy of solving one system by a Krylov subspace method and projecting residuals of other systems is related to the Lanczos-Gallerkin projection techniques discussed, for example, in [Parlett, LAA'80], [Saad, MC'87] and [Chan and Wan, SISC'97]. Building upon analysis of such techniques, we show that the effectiveness of the projections of a shifted residual associated to shift σ depends upon how large $|\sigma|$ is, and this analysis is backed up by numerical experiments.

A DEIM Induced CUR Factorization

Danny C. Sorensen, and Mark Embree

Abstract

I will present a CUR matrix factorization based upon the Discrete Empirical Interpolation Method (DEIM). A CUR factorization provides a low rank approximate factorization of a given matrix $\mathbf{A}$ of the form $\mathbf{A} \approx \mathbf{CUR}$ where $\mathbf{C}$ is a subset of the columns of $\mathbf{A}$ and $\mathbf{R}$ is a subset of the rows of $\mathbf{A}$. The matrix $\mathbf{U}$ is constructed so that $\mathbf{CUR}$ is a good approximation to $\mathbf{A}$. Assuming a low rank SVD $\mathbf{A} \approx \mathbf{VSW}^T$ is available, the DEIM points for $\mathbf{V}$ and $\mathbf{W}$ are used to select the matrices $\mathbf{C}$ and $\mathbf{R}$ respectively. This approximate factorization will satisfy $\|\mathbf{A} - \mathbf{CUR}\|_2 = \mathcal{O}(\sigma_{k+1})$, the first neglected singular value of $\mathbf{A}$ for a certain construction of $\mathbf{U}$.

An Efficient Algorithm for Computing the Generalized Null Space Decomposition

Nicola Guglielmi , Michael L. Overton and G. W. Stewart

Abstract

In an important paper on computing the Jordan form of a matrix Gene Golub and James Wilkinson [1] introduced a variant of the following decomposition of a matrix A of order n (for predecessors see [2, 3]). Specifically, if A is of index ν — i.e., if ν is the smallest integer for which $\text{null}(A^\nu) = \text{null}(A^{\nu+1})$ — then there is a unitary matrix V such that (for $\nu = 4$)

$$V^*AV = B = \begin{pmatrix} 0 & B_{12} & B_{13} & B_{14} & B_{15} \\ 0 & 0 & B_{23} & B_{24} & B_{25} \\ 0 & 0 & 0 & B_{34} & B_{35} \\ 0 & 0 & 0 & 0 & B_{45} \\ 0 & 0 & 0 & 0 & B_{55} \end{pmatrix}, \tag{1}$$

where

1. The leading diagonal blocks of B are square.
2. B_{12}, B_{23}, and B_{34} are of full column rank (which implies that, excluding B_{55}, the orders of the diagonal blocks of B are nonincreasing).
3. B_{55} is nonsingular (if it is not a 0×0 matrix).

If $V = (V_1 \ \cdots \ V_5)$, where the partitioning is conformal to that of B, then for $j = 1, \cdots, 4$ the column space of V_j is $\{x\colon A^j x = 0 \text{ and } A^{j-1}x \neq 0\}$, which we shall call the *generalized null space of A of degree j*. We will also call the decomposition the *generalized null space decomposition (GNSD)*. The dimensions of these subspaces completely determine the structure of the Jordan blocks for a zero eigenvalue of A.

Golub and Wilkinson gave a constructive proof of the existence of the decomposition. It begins with an orthonormal basis V_1 for the null space of A and uses it to deflate A into form

$$\begin{pmatrix} 0 & X \\ 0 & A' \end{pmatrix}. \tag{2}$$

The reduction continues by similarly deflating A' — and so on until the current matrix has no nontrivial null space. Golub and Wilkinson suggested using the SVD to determine the required null spaces. However, the number of SVD computations is the index ν. Hence if ν is near n (it equals n if A has a single Jordan block of order n), then the algorithm requires $O(n^4)$ operations.

In this talk we will describe an $O(n^3)$ algorithm for computing a GNSD. The procedure begins with a QR factorization and in its first step produces a QR factorization of A' in (2) by successively deflating approximate null vectors computed using a condition estimator. The tricky part is to do this in such a way that the estimator does not compute a generalized null vector of degree higher than one after the first deflation. The step concludes by applying a standard downdating algorithm to produce the QR factorization $A' = Q'R'$.[1] The algorithm proceeds as above by similarly reducing $Q'R'$. The procedure can be stably implemented using plane rotations.

[1]Note that although the downdating of an R factor alone is generally unstable, the simultaneous downdating of both Q and R factors is stable.

In either approach to computing the GNSD, the user must furnish a tolerance to determine when a computed singular value is zero or when an approximate null vector is satisfactory. Nonetheless, the computed GNSD can be shown to be the exact decomposition of a perturbation of A whose norm can be explicitly computed.

However, this property is not as useful as it appears. For if A is perturbed by rounding error, the result will in general be nonsingular and ν will be zero. Thus the proper question to ask about any algorithm for computing a GNSD is whether it can recover the generalized null space structure from the perturbed matrix. This is an open research problem, and we will discuss some of its aspects in the talk.

Finally, we will show how the GNSD of a matrix can be used to compute its Drazin generalized inverse. The advantage of this algorithm is that, aside from orthogonal transformations, it consists of solving a single Sylvester equation, which can be done efficiently using the output of our algorithm for computing the GNSD.

References

[1] G. H. Golub and J. H. Wilkinson. Ill-conditioned eigensystems and the computation of the Jordan canonical form. *SIAM Review*, 18:578–619, 1976.

[2] V. N. Kublanovskaja. On a method of solving the complete eigenvalue problem for a degenerate matrix. *USSR Computational Mathematics and Mathematical Physics*, 4:1–14, 1968. Originally appeared in Ž. Vyčisl. Math. Math. Fiz, 6:611–629, 1966.

[3] A. Ruhe. An algorithm for numerical determination of the structure of a general matrix. *BIT*, 10:196–216, 1970.

Fast Iterative Solvers for Fractional Differential Equations

Tobias Breiten[*], *Valeria Simoncini*[†] *and* *Martin Stoll*[‡]

Abstract

The study of fractional order differential equations is an old topic in mathematics going back to the likes of Euler and Leibniz. Despite its long history in mathematics it was not until recently that this topic has gained mainstream interest outside the mathematical community. This surging interest is mainly due to the inadequateness of traditional models to describe many real world phenomena. The well-known anomalous diffusion process is one typical such example. Other applications of fractional calculus are viscoelasticity, electrical circuits, electro-analytical chemistry or image processing.

In this talk we focus on a family of model problems that contain both the standard derivative as well as fractional derivatives of Caputo or Riemann-Liouville-type.

Consider a function $f(t)$ defined on an interval $[a, b]$. Assuming that $f(t)$ satisfies all requirements of a Caputo differentiable function of real order α with $(n-1 \leq \alpha < n)$, we obtain the Caputo derivative as

$$ {}_a^C D_t^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \int_a^t \frac{f^{(n)}(s)ds}{(t-s)^{\alpha-n+1}}. \tag{1} $$

Based on the discussion in [2] the Caputo derivative is frequently used for the derivative with respect to time. Additionally, we need to define the Riemann-Liouville derivative and again assuming that $f(t)$ satisfies the requirements for a left-sided fractional derivative of real order α with $(n-1 \leq \alpha < n)$,

$$ {}_a^{RL} D_t^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \left(\frac{d}{dt}\right)^n \int_a^t \frac{f(s)ds}{(t-s)^{\alpha-n+1}} \tag{2} $$

for $a < t < b$. Additionally, one is often very interested in evaluating right-side Riemann-Liouville fractional derivatives, which are given by

$$ {}_t^{RL} D_b^\alpha f(t) = \frac{(-1)^n}{\Gamma(n-\alpha)} \left(\frac{d}{dt}\right)^n \int_t^b \frac{f(s)ds}{(s-t)^{\alpha-n+1}} \tag{3} $$

for $a < t < b$. Additionally, we define the symmetric Riesz derivative

$$ \frac{d^\alpha f(t)}{d\,|t|^\alpha} =_t D_R^\alpha f(t) = \frac{1}{2}\left({}_a^{RL} D_t^\alpha f(t) + {}_t^{RL} D_b^\alpha f(t)\right). \tag{4} $$

A common way of discretizing differential equations of fractional order such as

$$ \frac{du(x,t)}{dt} - {}_x D_R^{\beta_1} u(x,t) = f(x,t), \tag{5} $$

where the fractional symmetric Riesz derivative is used, is to employ a Grünwald-Letnikov finite difference approximation. The resulting system matrix is then of Toeplitz form. This rather simple

[*]Institute of Mathematics and Scientific Computing, University of Graz, Heinrichstr. 36, A-8010 Graz, Austria, (tobias.breiten@uni-graz.at)

[†]Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, 40127 Bologna, Italy (valeria.simoncini@unibo.it)

[‡]Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg Germany, (stollm@mpi-magdeburg.mpg.de)

setup then requires the solution of a large scale linear system in Toeplitz form. Here, we introduce a suitable preconditioner of circulant-type that can be used within an iterative solver.

In the case that the equation now also contains a fractional Caputo time-derivative combined with the spatial Riesz derivative, i.e.,

$$ {}_0^C D_t^\alpha u(x,t) - {}_x D_R^\beta u(x,t) = f(x,t), \tag{6} $$

the individual discrete operators are again Toeplitz matrices but the overall structure of the linear system is of more involved form. In particular, using a space-time formulation a large-scale Sylvester equation needs to be solved. For this purpose, we here focus on two methods. The first method we discuss is the recently introduced KPIK method [3] and the second one is inspired by recent developments in the study of low-rank iterative methods for matrix equations [1, 4]. Namely, we focus on the use of a low-rank Krylov solver in combination with a preconditioner.

A similar structure within the linear systems is obtained when two- and three-dimensional problems are considered where each spatial and potentially the temporal domain is equipped with a fractional derivative. The resulting structured matrix equations are again solved using low-rank techniques or in case of a tensor structure we employ recently proposed tensor schemes.

References

[1] D. Kressner and C. Tobler, *Krylov subspace methods for linear systems with tensor product structure*, SIAM J. Matrix Anal. Appl, 31 (2010), pp. 1688–1714.

[2] I. Podlubny, A. Chechkin, T. Skovranek, Y. Chen, and B. Vinagre Jara, *Matrix approach to discrete fractional calculus II: Partial fractional differential equations*, Journal of Computational Physics, 228 (2009), pp. 3137–3153.

[3] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations.*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.

[4] M. Stoll and T. Breiten, *A low-rank in time approach to PDE-constrained optimization*, Submitted, (2013).

From PDEs through Functional Analysis to Iterative Methods, or there and back again

Josef Málek and Zdeněk Strakoš

Abstract

Consider the partial differential equation (PDE) problems described in the form of functional equation using the Hilbert space V and its dual $V^\#$

$$\mathcal{A}x = b, \quad \mathcal{A} : V \to V^\#, \quad x \in V, \quad b \in V^\#,$$

where the linear operator $\mathcal{A}$ is self-adjoint (with respect to the duality pairing $\langle \cdot, \cdot \rangle$), bounded and coercive. In solving PDE boundary-value problems (BVP), the state-of-the-art literature on using the conjugate gradient method (CG) (as well as of using other Krylov subspace methods) proceeds in most cases in the following way. First the linear algebraic system $\mathbf{Ax} = \mathbf{b}$ is formed by discretization, which is typically based on the bilinear form representation (so called weak formulation) and the sophisticated mathematical technology of the finite element method (FEM). Then the *standard unpreconditioned* CG is considered. Since in almost all cases this would result in a slow convergence of $\mathbf{x}_n$ to the algebraic solution $\mathbf{x}$, preconditioning of the discretized algebraic problem is introduced to accelerate the convergence behavior and save the game. *Preconditioned* CG is thus derived algebraically by preconditioning the discretized algebraic system and subsequently reformulating the computational algorithm in order to produce the approximation vectors and residuals corresponding to the original system. The solution process can be illustrated by the following schema:

$$a(\cdot,\cdot),\ \langle b, \cdot \rangle \ \to\ \mathbf{A}_h, \mathbf{b}_h \ \to\ \text{algebraic preconditioner} \ \to\ \text{PCG applied to } \mathbf{A}_h\mathbf{x}_h = \mathbf{b}_h\,,$$

where h stands for a discretization parameter. In this way, preconditioning is considered without being a part of the discretization of the infinite dimensional problem. The whole computational process for solving the original infinite dimensional problem, construction of efficient algorithms and their analysis is then broken into (more or less artificially separated) pieces, with a posteriori analysis of the discretization error in most times separated from considering inexact algebraic computations and evaluation of the algebraic error. This contradicts the basic principle of using iterative methods where the main advantage is stopping the iteration process whenever the appropriate accuracy has been reached.

If, however, the algebraic preconditioning is motivated by the so called *operator preconditioning*, then the solution process can be illustrated by the following schema:

$$\{\mathcal{A}, \tau\} \to \{\mathcal{A}_h, \tau\} \to \{\mathbf{A}_h, \mathbf{M}_h\} \to\ \text{PCG applied to } \mathbf{A}_h\mathbf{x}_h = \mathbf{b}_h\,,$$

where $\mathbf{M}_h$ is linked with the choice of the inner product $(\cdot, \cdot)$, i.e., it corresponds to the discrete representation of the Riesz map $\tau : V^\# \to V$ which transforms the original problem into

$$\tau\mathcal{A}x = \tau b, \quad \tau\mathcal{A} : V \to V, \quad x \in V, \quad \tau b \in V\,.$$

In this way, the operator formulation naturally leads to the *infinite dimensional* CG where the Riesz map τ represents preconditioning, and discretization immediately gives PCG. It should be noted that within this straightforward and elegant approach, *unpreconditioned* CG corresponds

to the very particular case where the discretization basis functions are mutually orthogonal with respect to the inner product $(\cdot, \cdot)$.

Operator preconditioning deals with operators and PDEs, and algebraic preconditioning deals with matrices. Discretization can be considered as a tool for getting from the infinite dimensional operator setting to the finite dimensional matrix algebraic setting (efficient algebraic preconditioning associated with the infinite dimensional operator preconditioning can be derived in some cases in a purely algebraic way). Following the practice in solving challenging problems, preconditioning and discretization should be linked together as much as possible. In particular, preconditioning can be interpreted as *transformation of the discretization basis*, which is a well known from, e.g., hierarchical basis preconditioning and some techniques in domain decomposition.

After recalling some previously published results and historical origin of operator preconditioning, we present simple general transformation formulas within the framework of the CG method and discuss possible implications.

Reduced Basis Method for Parameterized Lyapunov Equations

Nguyen Thanh Son and Tatjana Stykel

Abstract

Consider a parameter-dependent Lyapunov equation

$$A(p)X(p)E^T(p) + E(p)X(p)A^T(p) = -B(p)B^T(p), \tag{1}$$

where the matrices $E(p), A(p) \in \mathbb{R}^{n\times n}$, $B(p) \in \mathbb{R}^{n\times m}$ and the sought-after solution matrix $X(p) \in \mathbb{R}^{n\times n}$ depend on a parameter vector $p \in \mathbb{P} \subset \mathbb{R}^d$. It is assumed that $E(p)$ is symmetric, positive definite and $A(p)$ is symmetric, negative definite for all $p \in \mathbb{P}$. In this case, the Lyapunov equation (1) has a unique symmetric, positive semidefinite solution. Our goal is to solve the Lyapunov equation (1) on the whole parameter domain $\mathbb{P}$. Such a problem arises in balanced truncation model reduction of parametric linear control systems

$$\begin{aligned} E(p)\dot{x}(t,p) &= A(p)\,x(t,p) + B(p)u(t) \\ y(t,p) &= C(p)x(t,p) \end{aligned}$$

resulting, for example, from the finite element discretization of parameter-dependent parabolic PDEs, where parameters describe the geometry of a spatial domain and material properties.

In this talk, we discuss the numerical solution of the parametric Lyapunov equation (1) using a reduced basis method. This method was first proposed for model reduction of parametrically coercive linear elliptic PDEs and then extended to general noncoercive nonlinear parabolic problems, e.g., [1]. It has been proven to be the efficient and reliable solution method in a broad range of applications.

It is well known that the Lyapunov equation (1) can be written as a linear system

$$\boldsymbol{L}(p)\boldsymbol{x}(p) = \boldsymbol{b}(p), \tag{2}$$

with $\boldsymbol{L}(p) = -E(p)\otimes A(p) - A(p)\otimes E(p) \in \mathbb{R}^{n^2\times n^2}$, $\boldsymbol{x}(p) = \mathrm{vec}(X(p))$ and $\boldsymbol{b}(p) = \mathrm{vec}(B(p)B^T(p))$, where $\otimes$ denotes a Kronecker product and $\mathrm{vec}(\cdot)$ is a vector of length n^2 obtaining from the corresponding matrix by stacking its columns. The reduced basis method when applied to system (2) includes the following steps:

- snapshot collection: for selected parameters $p_1, \ldots, p_k \in \mathbb{P}$, construct a reduced basis matrix $\boldsymbol{V}_k = [\,\boldsymbol{x}(p_1),\, \ldots,\, \boldsymbol{x}(p_k)\,] \in \mathbb{R}^{n^2\times k}$, where $\boldsymbol{x}(p_j)$ is the solution of (2) with $p = p_j$, $j = 1, \ldots, k$;
- Galerkin projection: for any $p \in \mathbb{P}$, compute an approximate solution $\boldsymbol{x}(p) \approx \boldsymbol{V}_k\,\tilde{\boldsymbol{x}}(p)$, where $\tilde{\boldsymbol{x}}(p)$ solves the reduced system $\tilde{\boldsymbol{L}}(p)\tilde{\boldsymbol{x}}(p) = \tilde{\boldsymbol{b}}(p)$ with $\tilde{\boldsymbol{L}}(p) = \boldsymbol{V}_k^T\boldsymbol{L}(p)\boldsymbol{V}_k$ and $\tilde{\boldsymbol{b}}(p) = \boldsymbol{V}_k^T\boldsymbol{b}(p)$.

This seemingly simple procedure raises several issues that will be addressed in this talk: optimal choice of the sample parameters $p_1, \ldots, p_k$ providing good reduced basis spaces that guarantee a rapid convergence of the reduced basis approximation $\boldsymbol{V}_k\,\tilde{\boldsymbol{x}}(p)$ to $\boldsymbol{x}(p)$ over the entire parameter domain; conditioning of the reduced basis linear systems; error estimation for the approximate solution; computational cost and storage requirement.

If we assume that the solution of the Lyapunov equation (1) with the low-rank right-hand side $B(p)B^T(p)$ can be well approximated by a low-rank matrix, we can keep all operations in terms of $n \times n$ matrices and vectors of length n. Then the reduced basis method for the Lyapunov equation (1) is formulated as follows:

- snapshot collection: for selected parameters $p_1, \ldots, p_k \in \mathbb{P}$, construct a reduced basis matrix $V_k = [\, R_1, \ldots, R_k \,]$, where R_j is the low-rank Cholesky factor of the solution of (1) with $p = p_j$, $j = 1, \ldots, k$;
- Galerkin projection: for any $p \in \mathbb{P}$, compute an approximate solution $X(p) \approx V_k \tilde{X}(p) V_k^T$, where $\tilde{X}(p)$ solves the reduced Lyapunov equation

$$\tilde{A}(p)\tilde{X}(p)\tilde{E}^T(p) + \tilde{E}(p)\tilde{X}(p)\tilde{A}^T(p) = -\tilde{B}(p)\tilde{B}^T(p) \tag{3}$$

  with $\tilde{E}(p) = V_k^T E(p) V_k$, $\tilde{A}(p) = V_k^T A(p) V_k$ and $\tilde{B}(p) = V_k^T B(p)$.

For the construction of the reduced basis, we employ a Greedy algorithm that assures a small approximation error in the solution while keeping the dimension of the reduced basis as small as possible. For the success of this algorithm, a sharp and rigorous error estimate is required. Assuming an affine parameter dependence in the system matrices

$$E(p) = \sum_{j=1}^{n_E} \theta_j^E(p)\, E_j, \qquad A(p) = \sum_{j=1}^{n_A} \theta_j^A(p)\, A_j, \qquad B(p) = \sum_{j=1}^{n_B} \theta_j^B(p)\, B_j, \tag{4}$$

where E_j, A_j and B_j are parameter-independent, and $\theta_j^E, \theta_j^A, \theta_j^B : \mathbb{P} \to \mathbb{R}$, we use a min-$\theta$ approach [1] to derive an a posteriori error estimate for the approximate solution. This estimate is based on an error-residual relationship and can be computed efficiently using the affine decomposition (4). This decomposition can also be exploited to split the computation of the solution of (1) into an offline stage (computationally expensive), in which the reduced basis matrix V_k is constructed and the parameter-independent matrices $V_k^T E_j V_k$, $V_k^T A_j V_k$ and $V_k^T B_j$ are computed, and an online stage (computationally inexpensive), in which the reduced Lyapunov equation (3) is solved for any parameter $p \in \mathbb{P}$.

We also discuss the application of the reduced basis method for Lyapunov equations to parametric model order reduction by balanced truncation and present some results of numerical experiments.

References

[1] A.T. Patera, G. Rozza. *Reduced Basis Approximation and a Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2007.

Classical Iterative Methods for the Solution of Generalized Matrix Equations

Daniel B. Szyld

Abstract

There has been a flurry of activity in recent years in the area of solution of matrix equations. In particular, a good understanding has been reached on how to approach the solution of large scale (linear) Lyapunov equations and (nonlinear) Riccati equations; see, e.g., [1]–[5].

An effective way to solve both Lyapunov equations of the form $A^TX + XA + C^TC = 0$, where A and X are $n \times n$, or Riccati equations of the form $A^TX + XA - XBB^TX + C^TC = 0$, is to use Galerkin projection with appropriate extended or rational Krylov subspaces. These methods work in part because the solution is known to be symmetric positive definite with rapidly decreasing singular values, and therefore it can be approximated by a low rank matrix $X_k = Z_kZ_k^T$. Thus the computations are performed usually with storage which is lower rank, i.e., lower than order of n^2.

Generalized Lyapunov or Riccati equations have additional terms. In this paper, we concentrate on equations of the following form

$$A^TX + XA + \sum_{j=1}^{m} N_jXN_j^T + C^TC = 0,$$

$$A^TX + XA - XBB^TX + \sum_{j=1}^{m} N_jXN_j^T + C^TC = 0.$$

Such equations arise for example in stochastic control [6]. It has been proposed, e.g., in [7], to use approximations to conjugate gradients or to BiCGStab, appropriately preconditioned, where the basis vectors (matrices) and iterates are "truncated" throughout the algorithm to keep all these elements represented by low-rank matrices.

In the present work, we propose a return to classical iterative methods, and consider instead simple iterations, where the iterate $X_{k+1} = Z_{k+1}Z_{k+1}^T$ is the solution of $P(X) = Q(X_k)$, where $P(X)$ is A^TX+XA in the generalized Lyapunov case, and $A^TX+XA-XBB^TX$ in the generalized Riccati case; and $Q(X) = -\sum_{j=1}^{m} N_jXN_j^T - C^TC$; cf. [8] where such simple iterations were used to solve certain classes of Riccati equations. The classical theory of splittings applies here for the linear Lyapunov case, and the classical theory of contracting iterations for non-linear equations applies here for the Riccati case. This theory imposes some conditions on $P(X)$ and $Q(X)$.

One of the advantages of this classical approach is that only the data and the low-rank factors of the old and new iterates X_k and X_{k+1} need to be kept in storage. Furthermore, the solutions of the equations with $P(X)$, i.e., Lyapunov and Riccati, can be performed with the Galerkin projection methods mentioned above, where the growth of rank can usually be well contained. Numerical experiments show the competitiveness of the proposed approach.

This is joint work with Stephen D. Shank and Valeria Simoncini.

References

[1] Peter Benner, Jing-Rebecca Li, and Thilo Penzl. Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratric optimal control problems. *Numerical Linear Algebra with Applications*, 15:755–777, 2008.

[2] Vladimir Druskin, and Valeria Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. *Systems and Control Letters*, 60:546–560, 2011.

[3] Mohamed Heyouni and Khalid Jbilou. An extended block Arnoldi algorithm for large-scale solutions of the continuous-time algebraic Riccati equation. *Electronic Transactions on Numerical Analysis*, 33:53–62, 2009.

[4] Valeria Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM Journal on Scientific Computing*, 29:1268–1288, 2007.

[5] Valeria Simoncini, Daniel B. Szyld, and Marlliny Monsalve. On the numerical solution of large-scale Riccati equations. *IMA Journal on Numerical Analysis*. Available on line, 2013.

[6] Tobias Damm. *Rational Matrix Equations in Stochastic Control*. Lecture Notes in Control and Information Sciences, vol. 297, Springer, Berlin and Heidelberg, 2004.

[7] Peter Benner and Tobias Breiten. Low rank methods for a class of generalized Lyapunov equations and related issues. *Numerische Mathematik*, 124:441–470, 2013.

[8] Chun-Hua Guo and Alan J. Laub. On the iterative solution of a class of nonsymmetric algebraic Riccati equations. *SIAM Journal on Matrix Analysis and Applications*, 22:376–391, 2000.

Why does Shift-and-Invert Arnoldi work?

Christian Schröder and Leo Taslaman

Abstract

Virtually all methods for finding eigenvalues close to a target σ of a large matrix A involve solving linear systems with the shifted matrix $A - \sigma I$. If the shift is "too good," these systems become ill-conditioned or even singular, which at first sight seems inherently bad. Fortunately, at least for normal matrices, there is a strong connection between the largest possible forward errors of the linear systems and some of the eigenvectors in which we are interested. For certain algorithms, e.g., inverse iteration, this connection implies that the ill-conditioning is benign. For shift-and-invert Arnoldi, however, it is more complicated. We discuss this case using a backward error analysis. More precisely, we show how errors from solving linear systems, as well as from orthogonalizing vectors, lead to a perturbation of the underlying recurrence, and we bound the norm of this perturbation.

The standard Arnoldi algorithm (that is, without shift-and-invert) relies on the recurrence

$$AV_k = V_{k+1}\underline{H}_k, \tag{1}$$

where $V_{k+1} = [V_k, v_{k+1}]$ satisfies $V_{k+1}^T V_{k+1} = I$ and $\underline{H}_k$ is a $(k+1) \times k$ Hessenberg matrix. All theory on how the algorithm locates eigenvalues depends on this recurrence, but, due to floating point arithmetic, we do not see this recurrence in practice. Instead, what we have is

$$A\widehat{V}_k = \widehat{V}_{k+1}\underline{\widehat{H}}_k + E,$$

where $\widehat{V}_{k+1} = [\widehat{V}_k, \widehat{v}_{k+1}]$ and $\underline{\widehat{H}}_k$ are the computed quantities (so in general $\widehat{V}_{k+1}^T \widehat{V}_{k+1} \neq I$), and E is an error which is small relative to A. An important fact is that we can push E into a backward error with respect to A, and in this way recover an Arnoldi-like recurrence (called *Krylov decomposition* in [1]):

$$(A + \Delta A)\widehat{V}_k = \widehat{V}_{k+1}\underline{\widehat{H}}_k,$$

where $\Delta A = -E\widehat{V}_k^\dagger$ and $\widehat{V}_k^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\widehat{V}_k$ [2]. In this talk, we discuss the analog to this result for the shift-and-invert Arnoldi method, that is, when A in (1) is replaced by $(A - \sigma I)^{-1}$. We discuss which errors are being introduced, and then track these through the algorithm and push them back into a backward error ΔA with respect to A. We show that the computed quantities satisfy an exact recurrence

$$(A + \Delta A - \sigma I)^{-1}\widehat{V}_k = \widehat{V}_{k+1}\underline{\widehat{H}}_k,$$

and we bound $\|\Delta A\|$ in terms of $\|A\|$.

We give two bounds, which correspond to two different assumptions on the residuals of the linear systems that are being solved within the algorithm. The first bound assumes that the linear systems are solved in a backward stable fashion. This means that a computed solution $\widehat{x}$ of a linear system $Ax = b$ gives a residual that satisfies

$$\|A\widehat{x} - b\| \leq (\|A\|\|\widehat{x}\| + \|b\|)\epsilon,$$

where ϵ is a precision/tolerance parameter. If the computed basis $\widehat{V}_{k+1}$ is close to orthogonal, and the shift is meaningful (that is, $|\sigma| < \|A\|$), then we will see that the condition number of $\underline{\widehat{H}}_k$

dictates the norm of ΔA. The second bound assumes that the linear systems are solved with a residual that is small relative to the right hand side. In other words,

$$\|A\widehat{x} - b\| \leq \|b\|\epsilon.$$

This condition "may be very stringent, and possibly unsatisfiable"[3, p 336], but it is nevertheless often used in practice as a stopping condition for iterative linear solvers. Here, ϵ is often relatively large, say 10^{-6}. Our bound in this case can only be used if certain conditions (which can be cheaply verified during the computation) are true. However, when these conditions are satisfied, we get a particularly nice bound, independent of the condition number of $\underline{\widehat{H}}_k$.

Finally, we explain how these results can be extended to the case of generalized eigenvalue problems.

Our bounds are not only of importance to eigenvalue computation, but to any algorithm that relies on the shift-and-invert Arnoldi recurrence. Hence, other subfields of numerical analysis, for example, model order reduction and the computation of matrix functions, are also affected by our results.

References

[1] G.W. Stewart. Backward error bounds for approximate Krylov subspaces. Linear Algebra Appl. 340 81-86, 2002.

[2] G.W. Stewart. A Krylov–Schur algorithm for large eigenproblems. SIAM J. Matrix Anal. Appl. 23:3, 601–614, 2001.

[3] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

A new algorithm for computing quadrature-based bounds in CG

Gérard Meurant and Petr Tichý

Abstract

Today the (preconditioned) Conjugate Gradient (CG) algorithm by Hestenes and Stiefel is the iterative method of choice for solving linear systems $Ax = b$ with a real positive definite symmetric matrix A. An important question is when to stop the iterations. Ideally, one would like to stop the iterations when some norm of the error $x - x_k$, where x_k are the CG iterates, is small enough. However, the error is unknown and most CG implementations rely on stopping criteria that use the residual norm $\|r_k\| = \|b - Ax_k\|$ as a measure of convergence. These types of stopping criteria can provide misleading information about the actual error. It can stop the iterations too early when the norm of the error is still too large, or too late in which case too many floating point operations have been done for obtaining the required accuracy. This motivated researchers to look for ways to compute estimates of some norms of the error during CG iterations. The norm of the error which is particularly interesting for CG is the A-norm which is minimized at each iteration,

$$\|x - x_k\|_A \, = \, ((x - x_k)^T A(x - x_k))^{1/2}.$$

Inspired by the connection of CG with the Gauss quadrature rule for a Riemann-Stieltjes integral, a way of research on this topic was started by Gene Golub in the 1970s and continued throughout the years with several collaborators (e.g., Dahlquist, Eisenstat, Fischer, Meurant, Strakoš). The main idea of Golub and his collaborators was to obtain bounds for the integral using different quadrature rules. It turns out that these bounds can be computed without the knowledge of the stepwise constant measure and at almost no cost during the CG iterations.

These techniques were used by Golub and Meurant in 1994 for providing lower and upper bounds on quadratic forms $u^T f(A)u$ where f is a smooth function, A is a symmetric matrix and u is a given vector. Their algorithm GQL (Gauss Quadrature and Lanczos) was based on the Lanczos algorithm and on computing functions of Jacobi matrices. Later in [1], these techniques were adapted to CG to compute lower and upper bounds on the A-norm of the error for which the function is $f(\lambda) = \lambda^{-1}$. The idea was to use CG instead of the Lanczos algorithm, to compute explicitly the entries of the corresponding Jacobi matrices and their modifications from the CG coefficients, and then to use the same formulas as in GQL. The formulas were summarized in the CGQL algorithm (QL standing again for Quadrature and Lanczos), whose most recent version is described in the book [2].

The CGQL algorithm may seem complicated, particularly for computing bounds with the Gauss-Radau or Gauss-Lobatto quadrature rules. In this presentation based on our paper [4], we intend to show that the CGQL formulas can be considerably simplified. We use the fact that CG computes the Cholesky decomposition of the Jacobi matrix which is given implicitly, and derive new algebraic formulas by working with the LDL^T factorizations of the Jacobi matrices and their modifications instead of computing the Lanczos coefficients explicitly. In other words, one can obtain the bounds from the CG coefficients without computing the Lanczos coefficients. The algebraic derivation of the new formulas is more difficult than it was when using Jacobi matrices but, in the end, the formulas are simpler. Obtaining simple formulas is a prerequisite for analyzing the behaviour of the bounds in finite precision arithmetic and also for a better understanding of their dependence on the auxiliary parameters μ and η which are lower and upper bounds (or estimates) of the smallest and largest eigenvalues. We hope that these improvements will be useful for the implementation of quadrature-based error bounds into existing and forthcoming CG codes.

We first focus on explanation of the main idea of quadrature-based estimates of the A-norm of the error in CG. For simplicity, we will just concentrate on the Gauss and Gauss-Radau quadrature rules. Using the ideas of [3] and [5], we end up with the formula

$$\|x - x_k\|_A^2 \;=\; \widehat{Q}_{k,d} \;+\; \widehat{\mathcal{R}}_{k+d}\,,$$

where $\widehat{\mathcal{R}}_{k+d}$ stands for the remainder of the considered quadrature rule, $\widehat{Q}_{k,d}$ is computable, and $d > 0$ is a chosen integer. The remainder is positive when using the Gauss quadrature rule, and it is negative when using the Gauss-Radau quadrature rule with a prescribed node $\mu > 0$ that is strictly smaller than the smallest eigenvalue of A. Hence, $\widehat{Q}_{k,d}$ can provide a lower bound or an upper bound on $\|x - x_k\|_A^2$.

The question is how to compute $\widehat{Q}_{k,d}$ efficiently. The algorithm CGQL computes $\widehat{Q}_{k,d}$ using the entries of the corresponding Jacobi matrices and their rank-one or rank-two modifications. Our new algorithm CGQ (Conjugate Gradients and Quadrature) and its preconditioned version compute $\widehat{Q}_{k,d}$ directly from the CG coefficients. In particular, the lower bound based on the Gauss quadrature rule can be computed using the sum

$$\widehat{Q}_{k,d} = \sum_{j=k}^{k+d-1} \Delta_j, \qquad \Delta_j \equiv \gamma_j \|r_j\|^2,$$

where γ_j are the CG coefficients, see also [5], and the upper bound based on the Gauss-Radau quadrature rule can be computed using

$$\widehat{Q}_{k,d} = \sum_{j=k}^{k+d-2} \Delta_j + \Delta_k^{(\mu)}$$

where $\Delta_k^{(\mu)}$ is updated using the new formula

$$\Delta_k^{(\mu)} = \frac{\|r_k\|^2 \left(\Delta_{k-1}^{(\mu)} - \Delta_{k-1}\right)}{\mu \left(\Delta_{k-1}^{(\mu)} - \Delta_{k-1}\right) + \|r_k\|^2}\,, \qquad \Delta_0^{(\mu)} = \frac{\|r_0\|^2}{\mu}.$$

In the final numerical experiment we will illustrate some of the difficulties that may arise with modified quadrature rules when computing in finite precision arithmetic.

References

[1] G. H. Golub and G. Meurant, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[2] G. H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*, Princeton University Press, USA, 2010.

[3] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

[4] G. Meurant and P. Tichý, *On computing quadrature-based bounds for the A-norm of the error in conjugate gradients*, Numer. Algorithms, 62 (2013), pp. 163–191.

[5] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal., 13 (2002), pp. 56–80.

Exploiting Tropical Algebra in Numerical Linear Algebra

James Hook, Vanni Noferini, Meisam Sharify and Françoise Tisseur

Abstract

Tropical mathematics is the mathematics of the real numbers, together with the operations of addition and maximum (or minimum). Tropical algebra has been developed for modelling discrete event systems, for solving certain optimisation problems and as a tool for proving results in algebraic geometry. It allows the description, in a linear fashion, of phenomena that are nonlinear in the conventional algebra. Recent results in [4], [5], [7], and [8] indicate that tropical algebra can also be useful to numerical linear algebra. The objective of this talk is to explore diverse applications of tropical algebra to matrix computations.

The *tropical semiring* (or *max-plus semiring*) is denoted by $(\overline{\mathbb{R}}, \oplus, \otimes)$ with $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, and with the addition $\oplus$ and the multiplication $\otimes$ defined by

$$a \oplus b = \max(a, b), \quad a \otimes b = a + b$$

for all $a, b \in \overline{\mathbb{R}}$. Note $-\infty \oplus a = a \oplus -\infty = a$ and $0 \otimes a = a \otimes 0 = a$ so $-\infty$ and 0 are the zero and unit elements of the semiring, respectively.

Sharify [7] showed recently that the exponential of the *tropical roots* of the *tropical polynomial*

$$\mathsf{t}p(x) = \bigoplus_{k=0}^{d} \ln |a_k| \otimes x^{\otimes k} = \max_{0 \le k \le d} (\ln |a_k| + kx)$$

are good approximations of the roots of $p(x) = \sum_{k=0}^{d} a_k x^k$ as long as the tropical roots are well separated. What makes this result interesting is that the tropical roots of $\mathsf{t}p(x)$, which are the points of nondifferentiability of $\mathsf{t}p(x)$, require only $O(d)$ operations for their computation via the Newton polygon so they are cheaper to compute than the roots of the "classical" polynomial $p(x)$. The tropical roots of $\mathsf{t}p(x)$ have been used for a while in the package MPSolve (Multiprecision Polynomial Solver) [2] for the selection of the starting points in the Ehrlich-Aberth method. Sharify's result justifies why this selection is so effective.

Being able to cheaply locate the eigenvalues of $P(\lambda)$ is useful for the numerical solution of polynomial eigenvalue problems, in particular, when selecting the starting points in the Ehrlich-Aberth method [3], [4], or when choosing the contour for contour integral methods for large scale problems [1]. We show that Sharify's result extends to matrix polynomials

$$P(\lambda) = \lambda^d A_d + \cdots + \lambda A_1 + A_0 \in \mathbb{C}[\lambda]^{n \times n}, \quad A_d \neq 0$$

in the following way. To $P(\lambda)$ we associate the tropical scalar polynomial

$$\mathsf{T}p(x) = \bigoplus_{k=0}^{d} \ln(\|A_k\|) \otimes x^{\otimes k},$$

where $\|\cdot\|$ is any matrix norm subordinate to a vector norm. We identify sufficient conditions under which a tropical root α_j of $\mathsf{T}p(x)$ is a good approximation to the modulus of nk_j eigenvalues of $P(\lambda)$, where k_j is the multiplicity of α_j, thereby providing cheap localization results for the eigenvalues of P.

Our aim is to use the tropical roots of $\mathsf{T}p(x)$ to improve the numerical stability of polynomial eigensolvers based on linearizations of $P(\lambda)$ such as the MATLAB function `polyeig`. We note that currently none of these eigensolvers is backward stable for polynomials of degree 3 or higher. We use the tropical roots α_j of $\mathsf{T}p(x)$ to define eigenvalue parameter scalings of the form $\lambda = \alpha_j\mu$. We prove that `polyeig` combined with these scalings returns eigenpairs with small backward errors for eigenvalues, which in modulus are not too far from α_j. We also show that with such eigenvalue parameter scaling, the linearization process does not increase the eigenvalue condition numbers, again for those eigenvalues of P that are not too far from α_j in modulus [9].

Finally, we discuss the approximation properties of the *tropical eigenvalues* of $B = \big(\ln(|a_{ij}|)\big) \in \overline{\mathbb{R}}^{n\times n}$ as approximations to the eigenvalues of $A = (a_{ij}) \in \mathbb{C}^{n\times n}$. The tropical eigenvalues of B are the tropical roots of the permanent of $B \oplus z \otimes I$. We note that in tropical algebra, the permanent is no more difficult to compute than the determinant in the conventional algebra. The computation of the tropical eigenvalues of B is numerically stable and is cheaper than the computation of the eigenvalues of A, in particular when A is sparse [6].

References

[1] J. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. A numerical method for polynomial eigenvalue problems using contour integral. *Japan J. Indust. Appl. Math.*, 27(1):73–90, 2010.

[2] D. A. Bini and G. Fiorentino. Design, analysis, and implementation of a multiprecision polynomial rootfinder. *Numer. Algorithms*, 23(2-3):127–173, 2000.

[3] D. A. Bini and V. Noferini. Solving polynomial eigenvalue problems by means of the Ehrlich-Aberth method. *Linear Algebra Appl.*, 439(4):1130–1149, 2013.

[4] D. A. Bini, V. Noferini, and M. Sharify. Locating the eigenvalues of matrix polynomials. *To appear in SIAM J. Matrix Anal. Appl.*, 2013.

[5] S. Hammarling, C. J. Munro, and F. Tisseur. An algorithm for the complete solution of quadratic eigenvalue problems. *To appear in ACM Trans. Math. Software*, 2012.

[6] J. Hook and F. Tisseur. Tropical eigenvalues. MIMS EPrint, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2013. In preparation.

[7] M. Sharify. *Scaling Algorithms and Tropical Methods in Numerical Matrix Analysis: Application to the Optimal Assignment Problem and to the Accurate Computation of Eigenvalues.* PhD thesis, Ecole Polytechnique, Palaiseau, France, Sept. 2011.

[8] M. Sharify, V. Noferini, and F. Tisseur. Locating the eigenvalues of matrix polynomials: tropical versus Pellet. MIMS EPrint, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2013. In preparation.

[9] M. Sharify and F. Tisseur. Effect of tropical scaling on linearizations of matrix polynomials: backward error and conditioning. MIMS EPrint, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2013. In preparation.

On the Sensitivity of Matrix Functions to Random Noise

Serge Gratton, David Titley-Peloquin, Philippe Toint and Jean Tshimanga Ilunga

Abstract

How sensitive is a matrix function $F : \Omega \subset \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ to perturbations in its input? This is a fundamental question in numerical linear algebra. Sensitivity analyses are usually performed using condition numbers. The Frobenius-norm absolute condition number of F at A is defined as

$$\lim_{\delta\to 0} \sup_{\|H\|_F\leq\delta} \frac{\|F(A+H)-F(A)\|_F}{\delta}. \tag{1}$$

This is an asymptotic *worst-case* bound on $\|F(A+H)-F(A)\|_F/\|H\|_F$ in the *small perturbation* limit. Here we are interested in random perturbations. We attempt to quantify some of the properties of $\|F(A+H)-F(A)\|_F$, where F is Fréchet differentiable, $A \in \Omega$ is fixed, and the elements of H are random variables following various distributions. We investigate the following questions:

(a) How descriptive is the worst-case asymptotic analysis from (1) for small random perturbations?

(b) Can anything be said about the typical sensitivity of F even for large perturbations?

In the above, 'small' and 'large' refer to the norm of the covariance matrix of $\mathrm{vec}(H)$.

Specifically, let the elements of H be random variables such that $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$ for a given positive semi-definite matrix $\Sigma \in \mathbb{R}^{mn\times mn}$. Extending ideas of Fletcher [1] and Stewart [8], we bound for any $\tau > 0$

$$\limsup_{\sigma\to 0} \mathrm{Prob}\left\{\frac{\|F(A+H)-F(A)\|_F}{\sigma} \geq \tau\right\}.$$

The result holds independently of the distribution of the elements of H. It can be used to bound the median (or any other percentile) of the random variable $\|F(A+H)-F(A)\|_F/\sigma$ asymptotically in the limit as $\sigma \to 0$. It provides an answer to question (a) above.

Analogously to the condition number (1), the above holds for asymptotically small perturbations. For some important classes of functions F and for random matrices H whose elements follow some common distributions, using results of Sankar et. al. [7] we have also obtained bounds on

$$\mathrm{Prob}\Big\{\|F(A+H)-F(A)\|_F \geq \tau\Big\},$$

for any $\sigma > 0$ and $\tau > 0$. In contrast to the usual analysis based on first-order expansions and condition numbers, these bounds describe the sensitivity of F to *large* random perturbations. This provides at least a partial answer to question (b) above.

This stochastic analysis can be found in [2]. To illustrate we give examples from [3] involving the sensitivity of solutions to least squares problems and related regularization algorithms. Motivated by recent interest in the data assimilation and optimization communities in the sensitivity of Krylov subspace iterates (e.g., [9, 5, 6]), we also present a sensitivity analysis of these iterates to random perturbations based on our work in [4].

This abstract is based on joint work with Serge Gratton (ENSEEIHT-IRIT and CERFACS), Philippe Toint (Université de Namur), and Jean Tshimanga Ilunga (ENSEEIHT-IRIT).

References

[1] R. Fletcher. *Expected Conditioning*. IMA J. Numer. Anal., 5(3):247–273, 1985.

[2] S. Gratton and D. Titley-Peloquin. *Stochastic Conditioning of Matrix Functions*. CERFACS Technical Report, 2013.

[3] S. Gratton, D. Titley-Peloquin, and J. Tshimanga Ilunga. *Sensitivity and Conditioning of the Truncated TLS Solution*. SIAM J. Matrix Anal. Appl., 34(3):1257–1276, 2013.

[4] S. Gratton, D. Titley-Peloquin, P. Toint, and J. Tshimanga Ilunga. *Differentiating the Method of Conjugate Gradients*. SIAM J. Matrix Anal. Appl., to appear, 2013.

[5] A. M. Moore, H. G. Arango, and G. Broquet. *Estimates of Analysis and Forecast Errors Derived from the Adjoint of 4D-Var*. Mon. Weather Rev., 140(10):3183–3203, 2012.

[6] J. J. Moré and S. M. Wild. *Estimating Computational Noise*. SIAM J. Sci. Comput., 33(3):1292–1314, 2011.

[7] A. Sankar, D. Spielman, and S.-H. Teng. *Smoothed Analysis of the Condition Numbers and Growth Factors of Matrices*. SIAM J. Matrix Anal. Appl., 28(2):460–476, 2006.

[8] G. W. Stewart. *Stochastic Perturbation Theory*. SIAM Rev., 32(4):576–610, 1990.

[9] Y. Zhu and R. Gelaro. *Observation Sensitivity Calculations Using the Adjoint of the Gridpoint Statistical Interpolation (GSI) Analysis System.* Mon. Weather Rev., 136(1):335–351, 2008.

Enhancing Incomplete Cholesky Decompositions

Jiří Kopal, Jennifer Scott, Miroslav Tuma and Miroslav Rozložník

Abstract

Incomplete Cholesky decompositions represent an important component in the solution of large sparse symmetric positive-definite systems of equations. Such decompositions arise in a wide range of practical applications. Preconditioners based on the decomposition combined with a Krylov-space accelerator are routinely used in a number of production codes.

Many variants of the incomplete Cholesky decomposition and of its refinements have been proposed and used to solve practical problems. The enhancements of the basic procedure have varied from new mathematical ideas and algorithmic simplifications to more sophisticated implementations. If we try to systematize the decades of development, we could classify them as offering improvements either in the accuracy of the LL^T decomposition measured by a norm of the distance from the system matrix A or in the stability of the computed factors.

Neither of these goals can be separated from the way in which the Cholesky decomposition is implemented. Consider, for simplicity, sequential implementations; we could state that for small $k \geq 1$, a useful and robust decomposition can afford to spend k-times more time in the factorization than the simplest no-fill procedure. With this flexibility in mind, a space-efficient incomplete Cholesky factorization based on the concept of intermediate memory introduced by Tismenetsky was recently discussed in [2] (with the resulting software made available to the community [3]). As we observed from extensive numerical experiments, the use of intermediate memory can improve preconditioner accuracy. While a fixed bound on the memory used during the decomposition seems to be a practical must, another important challenge is to decide on the distribution of the entries in the factor columns. There have been a lot of important contributions on this that include, for example, exploiting structural information provided by the symbolic factorization of a complete factorization or using dropping rules based on estimated or computed inverses of the factors. In this talk, we introduce a new way to split the memory between that required for the factor and the intermediate memory used only in its construction. It is based on estimating norms of Schur complement updates and we consider it as a natural extension of the ideas presented in [2].

We also propose a new dropping approach. It is theoretically motivated and it extends the concept of a posteriori filtering from the approximate inverse decomposition in [1] to the incomplete Cholesky decomposition. The new approach drops entries dynamically such that the significance of factor entries in different columns is balanced. We believe that both our proposals may provide steps on the way to achieving more robust incomplete Cholesky decompositions.

This work was partially supported by the projects P201/13-06684S and 108/11/0853 of the Grant Agency of the Czech Republic and by EPSRC grant EP/I013067/1.

References

[1] M. Rozložník, J. Kopal and M. Tuma. Approximate inverse preconditioners with adaptive dropping, *submitted to International Journal of Advances in Engineering Software*, 2013.

[2] J. A. Scott and M. Tuma. On positive semidefinite modification schemes for incomplete Cholesky factorization. Technical Report RAL-TR-2013-P-005, 2013.

[3] J. A. Scott and M. Tuma. `HSL_MI28`: an efficient and robust limited memory incomplete Cholesky factorization code, *ACM Trans. Math. Software*, 2013, to appear

Convergence of Optimization Schemes on Sets of Low-rank Matrices and Tensors

André Uschmajew

Abstract

Low-rank optimization is classical in statistical sciences, but becomes more an more an important tool to tackle problems of high-dimension. The task takes the abstract form

$$\min_{x\in\mathcal{M}} J(x) \tag{1}$$

where $f\colon \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ is a real valued function of, say, real $n_1\times n_2\times\cdots\times n_d$ tensors (which can be huge), and $\mathcal{M}$ is some set of *low-rank tensors*. Examples for J include quadratic functions or the Rayleigh quotient. For tensors of order $d\geq 3$ there exist several concepts of rank, leading to different sets $\mathcal{M}$. The recently used notions of *hierarchical tensor rank* and *TT rank* result in sets $\mathcal{M}$ which are intersections of low-rank matrix sets. From simplicity let us therefore assume $\mathcal{M}$ is the set of $n_1\times n_2$ matrices with rank bounded by k.

There are two common approaches to generate a sequence $(x_n)\subseteq\mathcal{M}$ which hopefully converges to a solution of (1). In the first approach we make use of the fact that typically an explicit parametrization of the set $\mathcal{M}$ is explicitly known. It takes the form of a *multi-linear* map $\tau(u_1,u_2,\dots,u_p)$, for instance, for low-rank matrices,

$$\tau\colon \mathbb{R}^{n_1\times k}\times\mathbb{R}^{n_2\times k}\to\mathbb{R}^{n_1\times n_2},\ \tau(U,V)=UV^T.$$

Then the function $f=J\circ\tau$ is minimized using a block coordinate method without side conditions: given $x^n=\tau(u_1^n,u_2^n,\dots,u_p^n)$, choose a block $i^*\in\{1,2,\dots,p\}$, then calculate

$$u_{i^*}^{n+1}\in\operatorname*{argmin}_{u_{i^*}} f(u_1^n,\dots,u_{i^*-1}^n,u_{i^*},u_{i^*+1}^n,\dots,u_p^n),$$

and set

$$x^{n+1}=\tau(u_1^n,\dots,u_{i^*-1}^n,u_{i^*}^{n+1},u_{i^*+1}^n,\dots,u_p^n).$$

One major problem in analyzing the convergence of such a method is to bound the linear maps

$$\tau_{\mathbf{u}}^i(v_i)=\tau(u_1,\dots,u_{i-1},v_i,u_{i+1},\dots,u_p)$$

from *below*. This typically means to bound from below the k-th largest singular value of the *unfolding* (or *matricization*) of the tensor $\tau(u_1,u_2,\dots,u_p)$ into a matrix that represents the map $\tau_{\mathbf{u}}^i$. In a recent preprint [1] with Zhening Li (Portsmouth) and Shuzhong Zhang (Minneapolis) we were able to handle these obstacles for the problem of tensor rank-one approximation

$$\min_{u_1,u_2,\dots,u_d}\|x-u_1\otimes u_2\otimes\cdots\otimes u_d\|$$

when using a greedy block selection algorithm, and gave *global* convergence results which I would like to present.

A second approach to solve (1) are tangential projection methods on the set $\mathcal{M}$ itself. The use of Riemannian optimization techniques on matrix manifolds is well understood by now. They can be applied, for instance, to the manifold of matrices of *fixed* rank. However, a notorious problem

is that this manifold is not closed, thereby excluding global convergence results as they exist in nonlinear optimization, unless, again, one can control the behaviour of the smallest singular value. It would be therefore of interest to consider tangential projection methods on the set of bounded rank matrices, replacing the tangent space by the tangent cone at singular points where the rank drops. Such a method then would take the usual form

$$x_{n+1} = R_{x_n}(x_n + \alpha\xi_n),$$

with ξ_n being tangent vectors and R_{x_n} being a (usually nonsmooth) *retraction* from the tangent cone at x_n to the set $\mathcal{M}$. Most naturally, one can choose as retraction a best approximation in $\mathcal{M}$, which in the case of bounded rank matrices can be computed using the SVD (exploiting the low-rank structure of tangent vectors). The convergence analysis for the sequence (x_n) becomes non-standard at cluster points of lower rank. However, these singular points are not that bad as one might expect. The tangent cone at such a point is the disjoint union of limits of subspaces from the regular part and can hence be explicitly described. (The set of bounded rank matrices is still a nice object from the algebraic geometry view-point.) Using the language of convex optimization, I would like to explore the possibilities of how to extend global critical point convergence results for retracted line search algorithms based on the Łojasiewicz gradient inequality

$$|f(x) - f(x^*)|^{1-\theta} \leq \Lambda\|\nabla f(x)\|,$$

which is well know for analytic manifolds, to the set of bounded rank matrices and tensors. A crucial step here is to show that one can replace the gradient in the Lojasiewicz inequality by its projection on the negative tangent cone, and that the retraction on $\mathcal{M}$ is locally stable, even in singular points. This is ongoing research with Reinhold Schneider (Berlin).

References

[1] Z. Li, A. Uschmajew, S. Zhang, *On convergence of the maximum block improvement method*, Preprint 2013.

[4] T. Rohwedder, A. Uschmajew, *On local convergence of alternating schemes for optimization of convex problems in the tensor train format*, SIAM J. Numer. Anal. 51 (2), 1134-1162 (2013).

[4] A. Uschmajew, *Local convergence of the alternating least squares algorithm for canonical tensor approximation*, SIAM J. Matrix Anal. Appl. 33 (2), 639-652 (2012).

[4] A. Uschmajew, B. Vandereycken, *The geometry of algorithms using hierarchical tensors*, Linear Algebra Appl. 439 (1), 133-166 (2013).

Structured Data Fusion with Tensorlab

Laurent Sorber, Marc Van Barel and Lieven De Lathauwer

Abstract

Structured data fusion (SDF) is the practice of jointly factorizing one or more coupled data sets while optionally imposing structure on the factors. Each data set - stored as a dense, incomplete or sparse tensor - in a data fusion problem can be factorized with a different tensor decomposition. Currently, in Tensorlab, the user has the choice of the Canonical Polyadic Decomposition and Block Term Decomposition models. With a bit of effort it is easy to add new models as well. Structure can be imposed on the factors in a modular way and the user can choose from a library of predefined structures such as nonnegativity, orthogonality, Hankel, Toeplitz, Vandermonde, matrix inverse, and many more. By selecting the right structures, classical matrix decompositions such as the QR factorization, eigenvalue decomposition and singular value decomposition can be computed. Tensorlab uses a domain specific language (DSL) for modelling structured data fusion problems. The three key ingredients of an SDF model are (1) defining variables, (2) defining factors as transformed variables and (3) defining the data sets and which factors to use for their factorizations. Using different examples, the modelling capabilities of this DSL will be demonstrated and the effectiveness and efficiency of SDF with Tensorlab will be illustrated.

Compact Rational Krylov Methods for the Nonlinear Eigenvalue Problem

Roel Van Beeumen, Karl Meerbergen and Wim Michiels

Abstract

We present a new framework of Compact Rational Krylov (CORK) methods for solving the nonlinear eigenvalue problem (NLEP):

$$A(\lambda)x = 0,$$

where $\lambda \in \Omega \subseteq \mathbb{C}$ is called an eigenvalue, $x \in \mathbb{C}^n \setminus \{0\}$ the corresponding eigenvector, and $A : \Omega \to \mathbb{C}^{n\times n}$ is analytic on Ω. Linearizations are used for many years for solving polynomial eigenvalue problems [5]. The matrix polynomial $P(\lambda) = \sum_{i=0}^{d} \lambda^i P_i$, with $P_i \in \mathbb{C}^{n\times n}$, is transformed to a linear pencil $L(\lambda) = X - \lambda Y$, with $X, Y \in \mathbb{C}^{dn\times dn}$, so that there is a one-to-one correspondence between the eigenvalues of $P(\lambda)x = 0$ and $L(\lambda)u = 0$.

For the general nonlinear case, i.e., nonpolynomial eigenvalue problem, $A(\lambda)$ is first approximated by a matrix polynomial [1, 4, 7] or rational matrix polynomial [2] and then a convenient linearization is used. The linearizations used in the literature can all be written in the following form

$$L(\lambda) = \mathbf{A} - \lambda\mathbf{B},$$

where

$$\mathbf{A} = \left[\begin{array}{cccc} A_0 & A_1 & \cdots & A_{d-1} \\ \hline \multicolumn{4}{c}{M \otimes I_{n\times n}} \end{array}\right], \qquad \mathbf{B} = \left[\begin{array}{cccc} B_0 & B_1 & \cdots & B_{d-1} \\ \hline \multicolumn{4}{c}{N \otimes I_{n\times n}} \end{array}\right],$$

with $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{dn\times dn}$, $A_i, B_i \in \mathbb{C}^{n\times n}$, and $M, N \in \mathbb{C}^{(d-1)\times d}$. Note that the pencil $(\mathbf{A}, \mathbf{B})$ also covers the dynamically growing linearization pencils used in [3, 4, 7, 2]. The construction of the (rational) polynomial approximation of $A(\lambda)$ can be obtained using results on approximation theory or can be constructed dynamically during the solution process. The last two years, we developed various variants of the rational Krylov method based on these ideas [7, 2].

The major disadvantage of methods based on linearizations is the growing memory cost with the iteration count. It is possible to reduce this cost in specific cases, but in general, the memory cost is proportional to the degree of the polynomial. However, we can exploit the kronecker structure of the part below the first block row of the pencil $(\mathbf{A}, \mathbf{B})$ such that the memory cost is not proportional with the degree of the polynomial any more and only grows linearly with the iteration number. We therefore present the CORK family of rational Krylov methods, which use a generalization of the compact Arnoldi decomposition, proposed in [6]. All these methods construct a subspace $\mathbf{V} \in \mathbb{C}^{dn\times k}$, represented by two smaller matrices $Q \in \mathbb{C}^{n\times r}$ and $\mathbf{U} \in \mathbb{C}^{dr\times k}$ with orthonormal columns and $r \leq d + k$, such that

$$\mathbf{V} = (I_{d\times d} \otimes Q)\mathbf{U}.$$

In this way, the extra memory cost due to the linearization of the original eigenvalue problems is negligible.

For large-scale NLEPs, we present a two-level approach: at the first level, the large-scale NLEP is projected yielding a small (nonlinear) eigenvalue problem at the second level. Therefore, the method consists of two nested iterations. In the outer iteration, we construct an orthogonal basis Q and project $A(\lambda)$ or $L(\lambda)$ onto it. In the inner iteration, we solve the projected small linear eigenvalue problem

$$\widehat{L}(\lambda)u = 0, \qquad \widehat{L}(\lambda) = (I_{d\times d} \otimes Q)^* L(\lambda)(I_{d\times d} \otimes Q),$$

or the projected small NLEP

$$\widehat{A}(\lambda)x = 0, \qquad \widehat{A}(\lambda) = Q^* A(\lambda) Q.$$

The partial Schur decomposition of the linearization allows for efficient and reliable locking, purging and restarting. We translate these operations on the (approximate) invariant pairs of the underlying nonlinear eigenvalue problem, which leads to a two-level approach. We also illustrate the methods with numerical examples and give a number of scenarios where the methods performs very well.

References

[1] C. Effenberger and D. Kressner, *Chebyshev interpolation for nonlinear eigenvalue problems*, BIT, 52 (2012), pp. 933–951.

[2] S. Güttel, R. Van Beeumen, K. Meerbergen, and W. Michiels, *NLEIGS: A class of robust fully rational Krylov methods for nonlinear eigenvalue problems*, Tech. Report TW633, Department of Computer Science, KU Leuven, Leuven, 2013.

[3] E. Jarlebring, K. Meerbergen, and W. Michiels, *A Krylov Method for the Delay Eigenvalue Problem*, SIAM J. Sci. Comput., 32 (2010), pp. 3278–3300.

[4] E. Jarlebring, W. Michiels, and K. Meerbergen, *A linear eigenvalue algorithm for the nonlinear eigenvalue problem*, Numer. Math., 122 (2012), pp. 169–195.

[5] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Vector Spaces of Linearizations for Matrix Polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 971–1004.

[6] Y. Su, J. Zhang, and Z. Bai, *A compact Arnoldi algorithm for polynomial eigenvalue problems*, in Recent Advances in Numerical Methods for Eigenvalue Problems (RANMEP2008), Taiwan, Jan. 2008.

[7] R. Van Beeumen, K. Meerbergen, and W. Michiels, *A Rational Krylov Method Based on Hermite Interpolation for Nonlinear Eigenvalue Problems*, SIAM J. Sci. Comput., 35 (2013), pp. A327–A350.

Error Bounds and Aggressive Early Deflation for Extended QR Algorithms

Thomas Mach, Raf Vandebril and David Watkins

Abstract

The QR algorithm is a renowned method for computing all eigenvalues of an arbitrary small to medium sized matrix. A preliminary unitary similarity transformation to Hessenberg form is indispensable for keeping the computational complexity of the subsequent QR steps under control. When restraining computing time is the vital issue, we observe that the prominent role played by the Hessenberg matrix is sufficient but perhaps not necessary to fulfill this goal.

We will show, that a whole family of matrices, sharing the major qualities of Hessenberg matrices can be put forward. Instead of considering and operating directly on the matrices themselves, we will utilize their QR factorization, where the Q factor itself is decomposed in $n-1$ rotations $Q = G_{p(1)} \dots G_{p(n-1)}$, each G_i operating on rows i and $i+1$ and p a permutation of $\{1, \dots, n-1\}$. The order of the rotations in the decomposition of the Q factor encompasses several particular matrix types such as Hessenberg, inverse Hessenberg, and CMV matrices. In general these matrices exhibit an intriguing link with extended Krylov subspaces and as such they will be named extended Hessenberg matrices.

We will expand the classical QR algorithm to deal with generic extended Hessenberg matrices, without paying in computational complexity. It will be shown that the standard bulge chasing interpretation in the implicit QR versions nicely translates to a rotation chasing approach. Moreover, the numerical experiments illustrate that the ordering of the rotations has a significant impact on the number of iterations to be performed, e.g., in some particular cases, the number of iterations can be diminished strikingly.

We will also discuss the deflation criterion determining which rotations are close to the identity, allowing to split the original problem in parts. It will be shown that this deflation criterion can be considered to be optimal with respect to absolute and relative perturbations on the eigenvalues. Further, we present a generalization of aggressive early deflation to the extended QR algorithm, allowing the identification and deflation of already converged, but hidden, eigenvalues. Numerical results underpinning the power of aggressive early deflation are included.

Robust Integrators for the Dynamical Low-Rank Approximation using Rank-Structured Tensors

Christian Lubich, Ivan Oseledets and Bart Vandereycken

Abstract

Low rank tensors Given a d-dimensional tensor $X \in \mathbb{R}^{n\times n\times\cdots\times n}$, there a few ways to approximate X by a low-rank tensor such that the approximation needs considerably less parameters than n^d. In this talk, we consider approximation by tensor trains (TT) or matrix product states (MPS), and by hierarchical tensors (HT); see [GKT13] for a recent overview with applications. For example, a TT/MPS tensor of rank $\mathbf{k} = (k_0, k_1, \ldots, k_d)$ with $k_0 = k_d = 1$ is a tensor $Y \in \mathbb{R}^{n\times n\times\cdots\times n}$ that can be written element-wise as

$$Y(\ell_1, \ell_2, \ldots, \ell_d) = G_1(\ell_1) \cdot G_2(\ell_2) \cdots G_d(\ell_d),$$

where $G_i(\ell_i) \in \mathbb{R}^{k_{i-1}\times k_i}$ is a parameterization matrix of Y for each i and ℓ_i. The required parameters is thereby reduced from n^d to less than dnK^2 with $K := \max_i k_i$.

The TT/MPS/HT formats enjoy many useful properties that make them attractive numerically: they are stable, allow for explicit quasi-best approximations and are based on matrix algebra. In addition, it was recently shown in [HRS12, HOV13] that the sets

$$\mathcal{M}_{\mathbf{k}} = \{Y \in \mathbb{R}^{n\times n\times\cdots\times n} \colon \text{TT/MPS rank of } Y \text{ is } \mathbf{k}\}$$

are smooth embedded submanifolds of $\mathbb{R}^{n\times n\times\cdots\times n}$. The same holds for HT tensors; see [UV13].

Dynamical low rank Consider now the problem where $X(t)$ is a time-dependent d-dimensional tensor that satisfies a differential equation

$$\dot{X}(t) = F(X(t)), \qquad F \colon \mathbb{R}^{n\times n\times\cdots\times n} \to \mathbb{R}^{n\times n\times\cdots\times n}.$$

Instead of approximating $X(t)$ by some (quasi-best) $Y(t) \in \mathcal{M}_{\mathbf{k}}$ for each t separately,

$$Y(t) \in \mathcal{M}_{\mathbf{k}}, \qquad \|Y(t) - X(t)\| \to \min,$$

one can also consider the dynamical low-rank approximation of $Y(t)$,

$$\dot{Y}(t) \in T_{Y(t)}\mathcal{M}_{\mathbf{k}}, \qquad \|\dot{Y}(t) - F(Y(t))\| \to \min, \tag{1}$$

where $T_{\mathbf{Y}(t)}\mathcal{M}_{\mathbf{k}}$ is the tangent space of $\mathcal{M}_{\mathbf{k}}$ at $Y(t)$. For $\|\cdot\|$ the ℓ_2 norm, (1) becomes

$$\dot{Y}(t) = \mathrm{P}_{T_{Y(t)}\mathcal{M}_{\mathbf{k}}} F(Y(t)), \tag{2}$$

with $\mathrm{P}_{T_Y\mathcal{M}_{\mathbf{k}}}$ an orthogonal projection. This approach was introduced and analyzed for matrices in [KL07] and extended to TT/MPS/HT in [LRSV13].

Integrating a stiff ODE Since the dimension of $\mathcal{M}_{\mathbf{k}}$ is $O(ndK^2)$, one would like to avoid integrating (2) in the full space $\mathbb{R}^{n\times n\times\cdots\times n}$. The standard approach is to lift (1) to an equivalent ODE that evolves on the horizontal space of a certain principal fiber bundle. For example, in the case of matrices, we obtain

$$\dot{U} = \mathrm{P}_U^{\perp} F(USV^T)VS^{-1}, \quad \dot{S} = U^T F(USV^T)V, \quad \dot{V} = \mathrm{P}_V^{\perp} [F(USV^T)]^T US^{-T}, \tag{3}$$

where $Y(t) = U(t)S(t)V(t)^T$ is a time-dependent factorization of $Y(t)$ with orthonormal $U, V \in \mathbb{R}^{n\times k}$ and non-singular $S \in \mathbb{R}^{k\times k}$. Since the matrices U, V, S are small compared to an $n \times n$ matrix, integrating (3) can now in principle be done efficiently. The extension to TT/MPS and HT tensors is more technical but not difficult conceptually.

However, there is a fundamental numerical problem. Fix $\sigma_{\max}(Y(t)) = 1$. Observe that when $\sigma_{\min}(Y(t)) \to 0$, the r.h.s. of (3) will become ill-conditioned resulting in a severe step-size restriction for standard explicit integrators. This is very unfortunate since $\sigma_{\min}(X(t)) \to 0$ is bound to happen when we are aiming for increasingly good approximations of $Y(t)$ by $X(t)$.

Split projector time stepper The step-size restriction from above can be avoided by implicit time integration. This is however more involved than explicit integration and it will still require regularization of the S^{-1} terms. In the case of matrices, [LO13] proposed an explicit integrator based on a clever Lie–Trotter splitting of the projector $\mathrm{P}_{T_{\mathbf{Y}(t)}\mathcal{M}_{\mathbf{k}}}$. This integrator has remarkably robustness properties in the case of small singular values and it is considerably more accurate than standard explicit integrators.

In this talk, I will explain how this integrator can be extended to TT/MPS and HT using similar splitting ideas. While the extension to TT/MPS is fairly intuitive, it turns out to be more intricate for HT. Numerical experiments indicate that the new integrator is considerably faster and more accurate than integrating the formulation as a horizontal flow, like in (3).

References

[GKT13] L. Grasedyck, D. Kressner, and C. Tobler. A literature survey of low-rank tensor approximation technique. *GAMM-Mitteilungen*, 36(1):53—78, 2013.

[HOV13] J. Haegeman, T. J. Osborne, and F. Verstraete. Post-matrix product state methods: To tangent space and beyond. *Phys. Rev. B.*, 88(075133), 2013.

[HRS12] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. *Num. Math.*, 120(4):701–731, 2012.

[KL07] O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM J. Matrix Anal. Appl.*, 29(2):434–454, 2007.

[LO13] C. Lubich and I.V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT (in press)*, 2013.

[LRSV13] C. Lubich, T. Rohwedder, R. Schneider, and B. Vandereycken. Dynamical approximation of hierarchical Tucker and tensor-train tensors. *SIAM J. Matrix Anal. Appl.*, 34(2):470–494, 2013.

[UV13] A. Uschmajew and B. Vandereycken. The geometry of algorithms using hierarchical tensors. *Lin. Alg. Appl.*, 439:133—166, 2013.

The Anti-Triangular Factorization of Symmetric Matrices

Nicola Mastronardi and Paul Van Dooren

Abstract

Indefinite symmetric matrices occur in many applications, such as optimization, least squares problems, partial differential equations and variational problems. In these applications one is often interested in computing a factorization of the indefinite matrix that puts into evidence the inertia of the matrix or possibly provides an estimate of its eigenvalues. In this talk we present a new matrix decomposition that provides this information for any symmetric indefinite matrix by transforming it to a block anti-triangular form using orthogonal similarity transformations. We discuss several of the properties of the decomposition and show its use in the analysis of saddle point problems.

The Quest for a General Functional Tensor Framework for Blind Source Separation in Biomedical Data Processing

Sabine Van Huffel

Abstract

Substantial progress in data recording technologies and information processing in recent years have enabled acquisition and analysis of large amounts of biomedical data. Extraction of the underlying health relevant information patterns, called sources, is crucial for a reliable prediction of the underlying pathology or health condition. Examples are cardiac and respiratory rhythms, epileptic and neuronal spikes, QRS-complexes, bursts frequently occurring in Electroencephalograms (EEG), peaks emerging in Magnetic Resonance Spectroscopy (MRS), etc. Their extraction is a core problem in biomedical data processing. Blind source separation (BSS) consists in finding such source signals and the mixing mechanism, given only the raw signals, hence the term blind. Rapid advances in healthcare diagnostics and medical technologies open up new challenges for information processing.

The BIOTENSORS project, granted by the European Research Council (grant 339804, from Febr. 1,2014, till Febr. 1, 2019) intends to significantly push the state of the art in the important field of BSS. We will develop algebraically and numerically well-founded tensor techniques for BSS involving block terms, heterogeneous data sets and various constraints. We will use these techniques to face current grand challenges in biomedical data fusion. Indeed, in today's information society, more or less coupled data are everywhere (e.g. social network data) and the challenge is to find good ways to interpret them. We will however limit ourselves to concrete applications in biomedical data processing, in which the BIOMED research group, STADIUS division, KU Leuven, has built a strong expertise. The overall aim translates concretely into the following specific objectives.

1. Development of advanced algorithms for the computation of tensor decompositions, firmly based on numerical mathematics. These will also allow more general studies, involving block terms, coupled data sets and various types of constraints, relevant for biomedical applications.

2. Introduction of advanced approaches to BSS, based on Block Term Decompositions instead of Canonical Polyadic Decompositions, possibly involving multimodal data, possibly exploiting source structure, and allowing a broad set of constraints.

3. Development of reliable algorithms for tensor decomposition updating and tracking, with extensions to coupled data sets, which may be used in continuous patient monitoring.

4. Development of next generation BSS algorithms for specific biomedical areas. We focus on 3 applications for which extensive expertise and fully validated datasets are available, namely:

   - Metabolite quantification and brain tumour tissue typing using Magnetic Resonance Spectroscopic Imaging,
   - Functional monitoring including seizure detection and polysomnography,
   - Cognitive brain functioning and seizure zone localization using simultaneous Electroencephalography-functional MR Imaging integration.

5. Development of open-source BSS software for broad use.

By working directly at the forefront in close collaboration with the clinical scientists who actually use our software, we can have a huge impact. Starting from an overview of existing tensor based BSS algorithms in the above-mentioned areas, we explore the potential impact and evolutions of tensor based BSS for biomedical data processing. This is joint work with Lieven De Lathauwer.

Rank-Revealing Decompositions for Matrices with Multiple Symmetries

David Bindel, Charles Van Loan and Joseph Vokt

Abstract

Suppose an order-6 tensor $\mathcal{A}$ has the property that the value of $\mathcal{A}(i_1, i_2, i_3, j_1, j_2, j_3)$ does not change if the i-indices are permuted, or if the j-indices are permuted, or if the i-indices as a group are swapped with the j-indices as a group. Such a tensor has multiple symmetries and if it is smartly unfolded into a matrix A, then A itself has interesting structure above and beyond ordinary symmetry. In the case of the given example, there are permutation matrices Γ_1 and Γ_2 (both involving Kronecker products and perfect shuffles) such that both $\Gamma_1 A \Gamma_1^T$ and $\Gamma_2 A \Gamma_2^T$ equal A. We show how to compute a structure-preserving, low-rank approximation to A using LDL^T with diagonal pivoting together with a very cheap block diagonalization that is performed at the start. The full exploitation of structure has dramatic ramifications for efficiency and applications.

Two-level Methods with a Priori Chosen Convergence Factor

Panayot S. Vassilevski

Abstract

A two-level method for solving a linear system $Ax = b$ with a s.p.d. matrix, is characterized with a convergent smoother M, an interpolation matrix P and a coarse matrix $A_c = P^T AP$. The (inverse of the) two-level operator admits the following explicit form

$$B_{TL}^{-1} = \overline{M}^{-1} + (I - M^{-T}A)PA_c^{-1}P^T(I - AM^{-1}),$$

where $\overline{M} = M\left(M + M^T - A\right)^{-1} M^T$ is a symmetrization of M (i.e., if M stands for forward Gauss-Seidel, then $\overline{M}$ stands for symmetric Gauss–Seidel). The above formula is readily derived from the more familiar product iteration formula

$$E_{TL} \equiv (I - M^{-T}A)(I - PA_c^{-1}P^T A)(I - M^{-1}A),$$

which represents the composition of three processes; namely, the pre-smoothing iteration $I - M^{-1}A$, coarse-grid correction $I - PA_c^{-1}P^T A$, and post-smoothing iteration $I - M^{-T}A$. It is readily seen that B_{TL} is s.p.d., and that it provides A-convergent iteration (for any A-convergent smoother M). We have the estimates

$$\mathbf{v}^T A\mathbf{v} \leq \mathbf{v}^T B_{TL}\mathbf{v} \leq K_{TL}\mathbf{v}^T A\mathbf{v}. \tag{1}$$

The constant K_{TL}, as it is well-known, is related to the weak approximation property of the interpolation matrix P (or the coarse space Range (P)); namely, for any vector $\mathbf{v}$ there is a coarse interpolant $P\mathbf{v}_c$ such that

$$\|\mathbf{v} - P\mathbf{v}_c\|^2_{\widetilde{M}} \leq K_{TL}\ \mathbf{v}^T A\mathbf{v}. \tag{2}$$

Above, $\widetilde{M} = M^T(M + M^T - A)^{-1}M$ is the symmetrization of M^T. The best constant K_{TL} in (1) has been proven to admit the following characterization

$$K_{TL} = \max_{\mathbf{v}} \frac{\min_{\mathbf{v}_c} \|\mathbf{v} - P\mathbf{v}_c\|^2_{\widetilde{M}}}{\mathbf{v}^T A\mathbf{v}}.$$

In this talk, we investigate choices of P and M that lead to making K_{TL} as close to unity as desired in beforehand. Equivalently, the TL-convergence factor, $\varrho_{TL} = 1 - \frac{1}{K_{TL}}$, can be made as small as desired.

We apply our strategy to matrices A that can be assembled from local semi-definite ones $\{A_e\}$, that is, the global quadratic form $\mathbf{v}^T A\mathbf{v}$ is a sum of the local quadratic forms $\mathbf{v}_e^T A_e\mathbf{v}_e$, where $\mathbf{v}_e$ is the restriction of $\mathbf{v}$ to the local sets (of indices) e. Typical examples are finite element discretization matrices as well as graph Laplacian.

The procedure involves first, partitioning the set $\{e\}$ into non-overlapping ones (aggregates), solving eigenproblems for the local matrices associated with each such aggregate; selecting subset of eigenvectors in the lower part of the computed spectrum, and building a tentative interpolant $\overline{P}$ which by construction ensures the weak approximation property

$$\|\mathbf{v} - \overline{P}\mathbf{v}_c\|^2_D \leq \eta\ \mathbf{v}^T A\mathbf{v}.$$

D is a diagonal matrix, the so-called ℓ_1–smother for A. The actual P is obtained by smoothing-out $\overline{P}$ using appropriate matrix-polynomial

$$P = p_{\nu_a}(D^{-1}A)\overline{P}.$$

A first key step in the analysis is establishing energy boundedness of P of the form

$$\|\mathbf{v} - P\mathbf{v}_c\|_A^2 \leq (1+\tau)\,\|\mathbf{v}\|_A^2, \tag{3}$$

for an a priori chosen $\tau > 0$ as small as needed, whereas at the same time, we maintain weak approximation property of the form,

$$\|\mathbf{v} - P\mathbf{v}_c\|_D^2 \leq \eta\,\frac{1}{\tau}\,\mathbf{v}^T A\mathbf{v}.$$

The smoother M is also chosen as a matrix polynomial

$$M^{-1} = \Big(I - q_{\nu_s}(D^{-1}A)\Big)\,A^{-1}, \quad (q_{\nu_s}(0) = 1).$$

The next key step in the analysis is establishing a smoothing property of the following type:

$$\mathbf{v}^T\overline{M}\mathbf{v} \leq (1+\tau)\mathbf{v}^T A\mathbf{v} + \frac{1}{\tau}\frac{C}{\nu_s^2}\mathbf{v}^T D\mathbf{v}. \tag{4}$$

The constant $\tau > 0$ can be chosen a priori as small as needed, whereas the constant C is independent of τ. Combining the two major estimates (3)-(4), we get an estimate of the form (2) with a constant K_{TL} which can get as close to unity as needed by choosing τ sufficiently small and ν_s sufficiently large.

We illustrate the performance of the method on some PDE and non-PDE (graph Laplacian) examples.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

On the Relationship Between Nesterov's Optimal Convex Optimization Algorithm and Conjugate Gradient

Sahar Karimi and Stephen A. Vavasis

Abstract

Consider the problem of minimizing $f : \mathbf{R}^n \to \mathbf{R}$, where f is a C^1 strictly convex function. In particular, we assume that there exist two parameters $L, l > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{l}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Here and throughout this abstract, $\|\cdot\|$ refers to the Euclidean norm. Such functions have a unique minimizer which we will denote as $\mathbf{x}^*$. Classical steepest-descent guarantees a function value reduction of ϵ, i.e., $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \epsilon(f(\mathbf{x}^0) - f(\mathbf{x}^*))$, where $\mathbf{x}^k$ is the kth iterate, provided that $k \geq \text{const} \cdot \log(1/\epsilon)L/l$.

Nemirovsky and Yudin [2], Nesterov [3, 4] and subsequent authors have proposed 'optimal' algorithms for this problem, which have a bound of $k = \text{const} \cdot \log(1/\epsilon)\sqrt{L/l}$ iterations.

In the case that $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$, where A is a positive definite matrix, it is well known that $\mathbf{x}^* = A^{-1}\mathbf{b}$, that f is strictly convex with $L = \lambda_{\max}(A)$ and $l = \lambda_{\min}(A)$, and that conjugate gradient achieves a function value reduction of ϵ in $k = \text{const} \cdot \log(1/\epsilon)\sqrt{\kappa(A)}$ iterations. Thus, conjugate gradient is also 'optimal' in the sense of Nemirovsky and Yudin. In practice, conjugate gradient substantially outperforms the more general optimal algorithms when specialized to quadratic functions for two reasons: first, the constant is smaller for conjugate gradient; second, conjugate gradient is known to converge superlinearly.

In this work, we show that Nesterov's 1998 optimal algorithm can be hybridized with conjugate gradient to yield a method with the best properties of both methods, namely, the optimal complexity bound is retained, but in the case of quadratic functions (i.e., solving linear equations), the fast conjugate gradient convergence is replicated, assuming the line-search is exact.

A consequence is that the hybrid algorithm in practice can sometimes significantly outperform both Nesterov's method and also traditional nonlinear conjugate gradient methods (Fletcher-Reeves, Polak-Ribière, Hager-Zhang [1]) when applied to nonquadratic but strictly convex functions. Furthermore, additional insight into linear conjugate gradient is obtained.

References

[1] W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optimization*, 16:170–192, 2005.

[2] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley and Sons, Chichester, 1983. Translated by E. R. Dawson from *Slozhnost' Zadach i Effektivnost' Metodov Optimizatsii*, 1979, Glavnaya redaktsiya fiziko-matematicheskoi literatury, Izdatelstva "Nauka".

[3] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR (translated as Soviet Math. Dokl.)*, 269(3):543–547, 1983.

[4] Y. Nesterov. Introductory lectures on convex programming. Volume I: Basic course. Available on-line, 1998.

Numerical Methods for Computing the $\mathcal{H}_\infty$-Norm of Large-Scale Descriptor Systems

Peter Benner, Ryan Lowe and Matthias Voigt

Abstract

In this talk we consider linear time-invariant continuous-time descriptor systems of the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where E, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $x(t) \in \mathbb{R}^n$ is the descriptor vector, $u(t) \in \mathbb{R}^m$ is the input vector, and $y(t) \in \mathbb{R}^p$ is the output vector. Systems of this kind are the natural formulation of many dynamical models arising, e.g., in the simulation, control and optimization of electrical circuits, constrained multi-body systems, or certain semidiscretized PDEs.

Often it is very convenient to work with the transfer function, mapping inputs to outputs in the Laplace domain. Assuming $Ex(0) = 0$, the transfer function of (1) is given by

$$G(s) = C(sE - A)^{-1}B.$$

In many applications, for instance in robust control design, one is interested in the $\mathcal{H}_\infty$-norm of a stable and proper transfer function $G(\cdot)$, defined by

$$\|G\|_{\mathcal{H}_\infty} := \sup_{s\in\mathbb{C}^+} \|G(s)\|_2 = \sup_{\omega\in\mathbb{R}} \|G(\mathrm{i}\omega)\|_2 \,,$$

with $\mathbb{C}^+ := \{s \in \mathbb{C} : \mathrm{Re}(s) > 0\}$.

If the state-space dimension n of the system is small, there already exist methods to compute the $\mathcal{H}_\infty$-norm which are mostly based on computing the purely imaginary eigenvalues $\{\mathrm{i}\omega_1, \ldots, \mathrm{i}\omega_k\}$ of skew-Hamiltonian/Hamiltonian matrix pencils of the form

$$\lambda \begin{bmatrix} E & 0 \\ 0 & E^T \end{bmatrix} - \begin{bmatrix} A & \frac{1}{\gamma}BB^T \\ -\frac{1}{\gamma}C^TC & -A^T \end{bmatrix}. \tag{2}$$

Then we have $\|G(\mathrm{i}\omega_j)\|_2 = \gamma$ for $j = 1, \ldots, k$, and the frequency intervals with $\|G(\mathrm{i}\omega)\|_2 > \gamma$ can be easily computed. However, in the most recent form, this algorithm [1] uses a full structured factorization of (2) and thus cannot be applied if the state-space dimension is large. In this talk we present two approaches to overcome this problem.

The first method [2] exploits the relationship between the $\mathcal{H}_\infty$-norm and the structured complex stability radius of the transfer function $G(\cdot)$. This is the spectral norm of the smallest complex perturbation Δ that makes the perturbed transfer function

$$G_\Delta(s) := C(sE - (A + B\Delta C))^{-1}B \tag{3}$$

unstable, improper, or even not well-defined. The algorithm is based on *structured ε-pseudospectra* for $G(\cdot)$. This means that we consider all perturbations Δ with $\|\Delta\|_2 < \varepsilon$ and analyze how the poles of (3) might move. The set of all these perturbed poles is called the structured ε-pseudospectrum of $G(\cdot)$. To compute the structured complex stability radius we have to find the structured ε-pseudospectrum that touches the imaginary axis. This calculation is carried out

by a nested iteration. The inner iteration is adapted from [3] and computes the structured ε-pseudospectral abscissa for a *fixed value of* ε, that is the real part of the rightmost point in the structured ε-pseudospectrum. This iteration relies on the fact that the entire pseudospectrum can be realized by rank-1 perturbations, so an optimal perturbation can be efficiently computed. In the outer iteration, the value of ε is updated by Newton's method in order to drive the structured ε-pseudospectral abscissa to zero, similarly as in [4]. Since the inner iteration might only converge to a local maximizer, we discuss a way to obtain good initial values that indeed allows convergence to a global maximizer in most cases. These initial values are obtained by computing some *dominant poles* [5], i.e., those poles of $G(\cdot)$ that generate the largest local maxima of $\|G(\mathrm{i}\omega)\|_2$ for $\omega \in \mathbb{R}$.

The second method [6] that we present goes back to the algorithm from [1], but instead of the skew-Hamiltonian/Hamiltonian pencils (2) it uses closely related even pencils. Actually, we would have to compute *all* purely imaginary eigenvalues of these pencils to obtain *all* intervals for ω in which $\|G(\mathrm{i}\omega)\|_2 > \gamma$. However, we can relax this requirement when we make again use of the dominant poles of $G(\cdot)$. These are used to calculate shifts for a structure-preserving method for even eigenvalue problems [7] which computes eigenvalues close to these shifts. In this way we do not necessarily compute all frequency intervals with $\|G(\mathrm{i}\omega)\|_2 > \gamma$, but we can still obtain the frequency interval that contains the optimal frequency ω at which the $\mathcal{H}_\infty$-norm is attained.

In the talk we will compare both approaches with each other and also with [1, 4], and discuss advantages and disadvantages. Furthermore, we will present numerical results and demonstrate that both methods perform well, even for rather difficult examples.

References

[1] P. Benner, V. Sima, and M. Voigt. $\mathcal{L}_\infty$-norm computation for continuous-time descriptor systems using structured matrix pencils. *IEEE Trans. Automat. Control*, 57(1):233–238, 2012.

[2] P. Benner and M. Voigt. A structured pseudospectral method for $\mathcal{H}_\infty$-norm computation of large-scale descriptor systems. *Math. Control Signals Systems*, 2013. Accepted.

[3] N. Guglielmi and M. L. Overton. Fast algorithms for the approximation of the pseudospectral abscissa and pseudospectral radius of a matrix. *SIAM J. Matrix Anal. Appl.*, 32(4):1166–1192, 2011.

[4] N. Guglielmi, M. Gürbüzbalaban, and M. L. Overton. Fast approximation of the H_∞ norm via optimization of spectral value sets. *SIAM J. Matrix Anal. Appl.*, 34(2):709–737, 2013.

[5] J. Rommes and N. Martins. Efficient computation of multivariable transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.*, 21(4):1471–1487, 2006.

[6] R. Lowe and M. Voigt. $\mathcal{L}_\infty$-norm computation for large-scale descriptor systems using structured iterative eigensolvers. Preprint MPIMD/13-20, Max Planck Institute Magdeburg, Oct. 2013. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[7] V. Mehrmann, C. Schröder, and V. Simoncini. An implicitly-restarted Krylov subspace method for real symmetric/skew-symmetric eigenproblems. *Linear Algebra Appl.*, 436:4070–4087, 2012.

Antitriangular Factorization for Saddle Point matrices and the Null Space method

Jen Pestana and Andy Wathen

Abstract

Recently, Mastronadi and Van Dooren introduced the block antitriangular (or "Batman") factorization for symmetric indefinite matrices. This factorization employs only orthogonal similarity transforms and allows for the solution of linear equations in a straightforward way.

We will show the simplification of this algorithm for saddle point matrices and demonstrate that it represents the well-known Null Space method in precisely the same way that LU represents Gauss Elimination. The batman factorization has several attractive features that we will touch on: it is backwards stable, inertia revealing and is simply updated for augmented Lagrangian modifications. The relation to common preconditioning approaches will also be dicussed given time.

Fast, Stable, Computation of the Eigenvalues of Unitary-plus-rank-one Matrices

Jared L. Aurentz, Thomas Mach, Raf Vandebril and David S. Watkins

Abstract

We consider upper Hessenberg unitary-plus-rank-one matrices, that is, matrices of the form $A = \tilde{U} + \tilde{x}\tilde{y}^T$, where $\tilde{U}$ is unitary, and A is upper Hessenberg. This includes the class of Frobenius companion matrices, so methods for this type of matrix can be applied to the problem of finding the zeros of a polynomial.

The unitary-plus-rank-one structure is preserved by any method that performs unitary similarity transformations, including Francis's implicitly-shifted QR algorithm. We present a new implementation of Francis's algorithm that acts on a data structure that stores the matrix in $O(n)$ space and performs each iteration in $O(n)$ time.

We store A in the form $A = QR$, where Q is unitary and R is upper triangular. In this sense our method is similar to one proposed by Chandrasekaran, Gu, Xia, and Zhu (Oper. Theory Adv. Appl. 179 (2007), pp. 111–143), but our method stores R differently. Since Q is a unitary upper-Hessenberg matrix, it can be stored as a product $Q = Q_1 Q_2 \cdots Q_{n-1}$, where each Q_j is a Givens-like unitary tranformation that acts only on rows j and $j+1$. We call these Q_j *core transformations*. Both our algorithm and that of Chandrasekaran et. al. use this representation of Q. For R, they use a quasiseparable generator representation. Our representation scheme factors R in the form

$$R = C_{n-1} \cdots C_1 (B_1 \cdots B_{n-1} + e_1 y^T),$$

where the C_j and B_j are unitary core transformations. This is possible because R is also unitary-plus-rank-one.

Performing Francis iterations on an upper Hessenberg matrix A of the form

$$A = QR = Q_1 \cdots Q_{n-1} C_{n-1} \cdots C_1 (B_1 \cdots B_{n-1} + e_1 y^T)$$

is largely a matter of manipulating core transformations. We will show how to do this.

We will compare our algorithm to other fast (and slow) algorithms in terms of speed and accuracy.

Fast Hankel Tensor-Vector Products and Application to Exponential Data Fitting

Weiyang Ding, Yimin Wei and Liqun Qi

Abstract

This talk is contributed to a fast algorithm for Hankel tensor-vector products. For this purpose, we first discuss a special class of Hankel tensors that can be diagonalized by the Fourier matrix, which is called *anti-circulant* tensors. Then we obtain a fast algorithm for Hankel tensor-vector products by embedding a Hankel tensor into a larger anti-circulant tensor. The computational complexity is about $\mathcal{O}(m^2 n \log mn)$ for a square Hankel tensor of order m and dimension n, and the numerical examples also show the efficiency of this scheme. Moreover, the block version for multi-level block Hankel tensors is discussed as well. Finally, we apply the fast algorithm to exponential data fitting and the block version to 2D exponential data fitting for higher performance.

A CS Decomposition Method for Eigenvalues of Orthogonal Matrices

Daniela Calvetti, Lothar Reichel and Hongguo Xu

Abstract

A factorized condensed form of a real $2n\times 2n$ orthogonal matrix U is proposed. With the assumption that U does not take ± 1 as its eigenvalues, one is able to determine a real orthogonal matrix $\widetilde{Q}$ such that

$$\widetilde{Q}U\widetilde{Q}^T = \Sigma Z \Sigma Z^T,$$

where

$$\Sigma = \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix}, \quad Z = \begin{bmatrix} C_1 & S_2^T \\ S_1 & C_2^T \end{bmatrix},$$

Z is orthogonal and C_1, C_2, S_1, S_2 are all $n\times n$ upper bidiagonal. This factorization can be computed mainly with an improved version of the Schur parameter method proposed in [1].

Once the factorized form has been computed, a full CS decomposition algorithm is applied to compute the CS decomposition

$$\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} Z \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}^T = \begin{bmatrix} C & -S \\ S & C \end{bmatrix},$$

where U_1, U_2, V_1, V_2 are real orthogonal and C, S are diagonal satisfying $C^2 + S^2 = I_n$. The proposed CS decomposition algorithm adopts the idea used in Demmel-Kahan's SVD algorithm ([2]) for computing the small CS values and uses an improved version of the method in [3, 4]. for the rest. Therefore, the computed CS values have high accuracy.

Define $Q = \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} \widetilde{Q}$. One has

$$QUQ^T = \begin{bmatrix} C^2 - S^2 & 2CS \\ -2CS & C^2 - S^2 \end{bmatrix},$$

which is essentially a Schur decomposition of U.

Because the CS iterations are performed on half-sized blocks, the method is less expensive than other QR type algorithms when the similarity matrix Q is needed.

References

[1] A. Bunse-Gerstner and L. Elsner, Schur parameter pencils for the solution of the unitary eigenproblem, *Linear Algebra Appl.*, 154/156:741Ð778, 1991.

[2] J. Demmel and W. Kahan, Accurate singular value of bidiagonal matrices, *SIAM J. Sci. Stat. Comput.* 11:873-912, 1990.

[3] B. D. Sutton, Computing the complete CS decomposition, *Numer. Algor.* 50:33-65, 2009.

[4] B. D. Sutton, Stable computation of the CS decomposition: simultaneous bidiagonalization, *SIAM J. Matrix Anal. Appl.* 33:1-21, 2012.

The Dual Padé Family of Iterations for the Matrix p-Sector Function and one topic more on a Specific Procrustes Problem

Krystyna Ziętak

Abstract

In the talk we focus on the dual Padé families of iterations, introduced in [15], for computing the matrix pth root and the matrix p-sector function. We determine certain regions of convergence of these iterations using results proved in [5] and [6]. The matrix p-sector function, introduced in [14], is a generalization of the matrix sign function.

The second topic of our talk is focused on a specific Procrustes problem. In a Procrustes problem the goal is to find a matrix X from some class $\mathbb{M}$ of matrices for which the following minimum is reached

$$\min_{X\in\mathbb{M}} \|A - BX\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm and A, B are given, real matrices. It is a nearness problem (for other nearness problems see [10]). In the classical orthogonal Procrustes problem, $\mathbb{M}$ is the set of orthogonal matrices. However, one considers also other types of $\mathbb{M}$, for example, the Stiefel matrices, symmetric matrices and so on (see [1], [3], [4], [9], [11, Theorem 8.6]). In our talk the set $\mathbb{M}$ contains matrices which are subblocks of orthogonal matrices. We investigate this specific Procrustes problem and present a preliminary version of an iterative algorithm for computing its solutions. This part of our talk is based on a joint work with João Cardoso (Coimbra, Portugal).

The dual Padé family of iterations for computing the matrix pth root includes the Schröder iterations, considered in [2], and the Newton and Halley methods. Therefore our results on a convergence of the dual Padé iterations generalize the results of Cardoso and Loureiro [2], Guo [8], Lin [13] for the Schröder iterations, the Newton and Halley methods, respectively. It is worth to underline that the Newton method does not belong to the Padé family of iterations introduced in [12]. However, the Halley method belongs to the Padé and dual Padé families of iterations, simultaneously.

The concept of the dual Padé family for computing the p-sector function is different from the idea of the reciprocal Padé family of iterations introduced by Greco, Iannazzo and Poloni [7] for the matrix sign function.

Guo [8] and Lin [13] have shown some properties of iterates and of corresponding residuals generated by the Newton and Halley methods for computing the principal pth root of a matrix. We generalize their results to the Padé family of iterations and to the dual Padé family of iterations for the matrix pth root. We prove that the properties of the Newton and Halley methods, presented by Guo [8], follow from properties of the Padé approximants to the function $(1-z)^{-1/p}$ that generate the iterations of the Padé and dual Padé families. For all above purposes, concerning the Padé and dual Padé families, we study the properties of Padé approximants to the function $(1-z)^{-1/p}$.

References

[1] L.-E. Andersson, T. Elfving, A constrained Procrustes problem, SIAM J. Matrix Anal. Appl. 18 (1997), 124–139.

[2] J.R. Cardoso, A.F. Loureiro, On the convergence of Schröder iteration functions for pth roots of complex numbers, Appl. Math. Comput. 217 (2011), 8833–8839.

[3] M.T. Chu, N.T. Trendafilov, The orthogonally constrained regression revisited, J. Comput. Graph. Stat. 10 (2001), 746–771.
[4] L. Eldén, H. Park, A Procrustes problem on the Stiefel manifold, Numerische Math. 82 (1999), 599-619.
[5] O. Gomilko, F. Greco, K. Ziętak, A Padé family of iterations for the matrix sign function and related problems, Numerical Linear Alg. Appl. 19 (2012), 585-605.
[6] O. Gomilko, D.B. Karp, M. Lin, Ziętak, Regions of convergence of a Padé family of iterations for the matrix sector function and the matrix pth root, J. Comput. Appl. Math. 236 (2012), 4410–4420.
[7] F. Greco, B. Iannazzo, F. Poloni, The Padé iterations for the matrix sign function and their reciprocals are optimal, Linear Alg. Appl. 436 (2012), 472–477.
[8] Ch.-H. Guo, On Newton's method and Halley's method for the principal pth root of a matrix, Linear Alg. Appl. 432 (2010), 1905–1922.
[9] N.J. Higham, The symmetric Procrustes problem, BIT 28 (1988), 133–143.
[10] N.J. Higham, Nearness Problems in Numerical Linear Algebra, PhD thesis, University of Manchester, Manchester 1985.
[11] N.J. Higham, Functions of Matrices. Theory and Computation, SIAM, Philadephia 2008.
[12] B. Laszkiewicz, K. Ziętak, A Padé family of iterations for the matrix sector function and the matrix pth root, Numerical Linear Alg. Appl. 16 (2009), 951–970.
[13] M. Lin, A residual recurrence for Halley's method for the matrix pth root, Linear Alg. Appl. 432 (2010), 2927–2930.
[14] L.S. Shieh, Y.T. Tsay, C.T. Wang, Matrix sector functions and their applications to system theory, IEEE Proc. 131 (1984), 171–181.
[15] K. Ziętak, The dual Padé families of iterations for the matrix pth root and the matrix p-sector function, J. Comput. Appl. Math., in press, online: 6 August 2013.

List of speakers

Page	Speakers	Misc
7	Absil Pierre-Antoine	U.C.Louvain Louvain-la-Neuve - Belgium absil@inma.ucl.ac.be
9	Aishima Kensuke	The University of Tokyo Bunkyo ku - Japan Kensuke_Aishima@mist.i.u-tokyo.ac.jp
11	Al-Mohy Awad	King Khalid University Abha - Saudi Arabia aalmohy@hotmail.com
13	Antoulas Athanasios	Rice University Houston - United States aca@rice.edu
16	Arioli Mario	STFC Didcot - United Kingdom mario.arioli@stfc.ac.uk
18	Avron Haim	IBM Research Yorktown Heights - USA haimav@us.ibm.com
20	Bai Zhaojun	University of California Davis - USA bai@cs.ucdavis.edu
22	Ballard Grey	Sandia National Laboratories Livermore - United States gmballa@sandia.gov
24	Barlow Jesse	The Pennsylvania State University University Park - United States barlow@cse.psu.edu
25	Beattie Christopher	Virginia Tech Blacksburg - United States beattie@vt.edu
26	Benner Peter	Max Planck Institute Magdeburg - Germany benner@mpi-magdeburg.mpg.de
28	Benzi Michele	Emory University Atlanta - USA benzi@mathcs.emory.edu
29	Bientinesi Paolo	RWTH Aachen Aachen - Germany pauldj@aices.rwth-aachen.de
31	Bindel David	Cornell University Ithaca - United States bindel@cs.cornell.edu

continued ...

. . . continued

Page	Speakers	Misc
32	Bolten Matthias	University of Wuppertal Wuppertal - Germany bolten@math.uni-wuppertal.de
34	Bora Shreemayee	Indian Institute of Technology Guwahati Guwahati - India shbora@iitg.ac.in
36	Boumal Nicolas	UCLouvain Louvain-la-Neuve - Belgium nicolasboumal@gmail.com
38	Boutsidis Christos	IBM Yorktown Heights - United States christos.boutsidis@gmail.com
39	Carden Russell	University of Kentucky Lexington - USA russell.l.carden@uky.edu
40	Carson Erin	University of California, Berkeley Berkeley - USA ecc2z@eecs.berkeley.edu
42	Chaturantabut Saifon	Thammasat Univeristy Pathumthani - Thailand schaturantabut@gmail.com
44	Chen Jie	Argonne National Laboratory Lemont - USA jiechen@mcs.anl.gov
45	Chow Edmond	Georgia Institute of Technology Atlanta - USA echow@cc.gatech.edu
47	Chung Julianne	Virginia Tech Blacksburg - USA jmchung@vt.edu
49	Deadman Edvin	University of Manchester Manchester - United Kingdom edvin.deadman@manchester.ac.uk
51	De Lathauwer Lieven	KU Leuven Kortrijk - Belgium Lieven.DeLathauwer@kuleuven-kulak.be
52	Demmel James	University of California, Berkeley Berkeley - USA demmel@berkeley.edu
53	de Sturler Eric	Virginia Tech Blacksburg - USA sturler@vt.edu

continued . . .

. . . continued

Page	Speakers	Misc
56	De Teran Fernando	Universidad Carlos III de Madrid Leganés - Spain fteran@math.uc3m.es
57	Dhillon Inderjit	The University of Texas at Austin Austin - United States inderjit@cs.utexas.edu
58	Dmytryshyn Andrii	Umeå University Umeå - Sweden andrii@cs.umu.se
60	Dopico Froilan	Universidad Carlos III de Madrid Leganés - Spain dopico@math.uc3m.es
61	Drineas Petros	Rensselaer Polytechnic Institute Troy - USA drinep@cs.rpi.edu
63	Drmac Zlatko	University of Zagreb, Faculty of Science Zagreb - Croatia drmac@math.hr
64	Druskin Vladimir	Schlumberger Doll Research Cambridge - United States druskin1@slb.com
65	Duff Iain	STFC,RAL and CERFACS Didcot - UK iain.duff@stfc.ac.uk
68	Duintjer Tebbens Jurjen	Academy of Sciences of the Czech Republic Praha 8 , Liben - Czech Republic duintjertebbens@cs.cas.cz
70	Edelman Alan	MIT Cambridge - United States edelman@mit.edu
71	Eldén Lars	Linköping University Linköping - Sweden lars.elden@liu.se
72	Elman Howard	University of Maryland College Park - USA elman@cs.umd.edu
73	Embree Mark	Rice University Houston - USA embree@rice.edu
75	Faßbender Heike	TU Braunschweig Braunschweig - Germany h.fassbender@tu-bs.de

continued . . .

. . . continued

Page	Speakers	Misc
77	Freitag Melina	University of Bath Bath - United Kingdom m.freitag@maths.bath.ac.uk
79	Frommer Andreas	Bergische Universität Wuppertal Wuppertal - Germany frommer@math.uni-wuppertal.de
81	Gander Martin	University of Geneva Geneva - Switzerland martin.gander@unige.ch
82	Gazzola Silvia	University of Padua Padova - Italy gazzola@math.unipd.it
84	Ghysels Pieter	Lawrence Berkeley National Lab Berkeley - United States pghysels@lbl.gov
86	Gillis Nicolas	Universite de Mons Mons - Belgium nicolas.gillis@umons.ac.be
87	Giraud Luc	INRIA Toulouse - FRANCE luc.giraud@inria.fr
88	Greenbaum Anne	University of Washington Seattle - USA greenbau@amath.washington.edu
89	Greif Chen	The University of British Columbia Vancouver - Canada greif@cs.ubc.ca
91	Grigori Laura	INRIA Paris - France Laura.Grigori@inria.fr
93	Grubišić Luka	University of Zagreb Zagreb - Croatia luka.grubisic@math.hr
95	Güttel Stefan	The University of Manchester Manchester - United Kingdom stefan.guettel@manchester.ac.uk
97	Gugercin Serkan	Virginia Tech. Blacksburg - USA gugercin@math.vt.edu
99	Guo Chun-Hua	University of Regina Regina - Canada chun-hua.guo@uregina.ca

continued . . .

. . . continued

Page	Speakers	Misc
101	Gutknecht Martin	ETH Zurich Zurich - Switzerland mhg@math.ethz.ch
102	Hanke Martin	University of Mainz Mainz - Germany hanke@math.uni-mainz.de
104	Hansen Per Christian	Technical University of Denmark Kgs. Lyngby - Denmark pcha@dtu.dk
105	Higham Nicholas	The University of Manchester Manchester - UK nick.higham@manchester.ac.uk
107	Hnětynková Iveta	Charles University in Prague Prague 8 - Czech Republic hnetynkova@cs.cas.cz
109	Hochstenbach Michiel	TU Eindhoven Eindhoven - The Netherlands m.e.hochstenbach@tue.nl
110	Iannazzo Bruno	Università degli Studi di Perugia Perugia - Italy bruno.iannazzo@dmi.unipg.it
112	Ipsen Ilse	North Carolina State University Raleigh - USA ipsen@ncsu.edu
113	Jarlebring Elias	KTH , Royal Institute of Technology Stockholm - Sweden eliasj@kth.se
115	Jeuris Ben	KU Leuven Leuven - Belgium ben.jeuris@cs.kuleuven.be
116	Kågström Bo Kågström	Umeå University Umeå - Sweden bokg@cs.umu.se
118	Khabou Amal	The University of Manchester Manchester - UK amal.khabou@manchester.ac.uk
120	Kilmer Misha	Tufts University Medford - USA misha.kilmer@tufts.edu
122	Knyazev Andrew	Mitsubishi Electric Research Laboratories Cambridge - USA knyazev@merl.com

continued . . .

. . . continued

Page	Speakers	Misc
123	Koskela Antti	University of Innsbruck Innsbruck - Austria antti.koskela@uibk.ac.at
125	Kressner Daniel	EPF Lausanne Lausanne - Switzerland daniel.kressner@epfl.ch
127	Kürschner Patrick	Max Planck Institute Magdeburg - Germany kuerschner@mpi-magdeburg.mpg.de
129	Langou Julien	University of Colorado Denver Denver - USA julien.langou@ucdenver.edu
131	Le Borne Sabine	Hamburg University of Technology Hamburg - Germany leborne@tuhh.de
133	Liesen Jörg	TU Berlin Berlin - Germany liesen@math.tu-berlin.de
135	Lin Lijing	The University of Manchester Manchester - United Kingdom lijing.lin@manchester.ac.uk
137	Li Ren-Cang	University of Texas at Arlington Arlington - USA rcli@uta.edu
138	Li Shengguo	National University of Defense Technology Changsha - China nudtlsg@gmail.com
139	Li Xiaoye	Lawrence Berkeley National Laboratory Berkeley - USA xsli@lbl.gov
141	Mackey D. Steven	Western Michigan University Kalamazoo - USA steve.mackey@wmich.edu
143	Mahoney Michael	University of California, Berkeley Berkeley - USA mahoneymw@gmail.com
144	Markovsky Ivan	Vrije Universiteit Brussel Brussels - Belgium ivan.markovsky@vub.ac.be
146	Mastronardi Nicola	Consiglio Nazionale delle Ricerche Bari - Italy n.mastronardi@ba.iac.cnr.it

continued . . .

. . . continued

Page	Speakers	Misc
147	Meerbergen Karl	KU Leuven Leuven - Belgium Karl.Meerbergen@cs.kuleuven.be
149	Mehl Christian	TU Berlin Berlin - Germany mehl@math.tu-berlin.de
151	Mehrmann Volker	TU Berlin Berlin - Germany mehrmann@math.tu-berlin.de
152	Meini Beatrice	University of Pisa Pisa - Italy meini@dm.unipi.it
154	Mengi Emre	Koc University Istanbul - Turkey emengi@ku.edu.tr
156	Meurant Gerard	CEA Paris - France gerard.meurant@gmail.com
158	Międlar Agnieszka	TU Berlin Berlin - Germany miedlar@math.tu-berlin.de
160	Moler Cleve	MathWorks, Inc Santa Fe - USA moler@mathworks.com
161	Morgan Ron	Baylor University Waco - USA Ronald_Morgan@baylor.edu
163	Nabben Reinhard	TU Berlin Berlin - Germany nabben@math.tu-berlin.de
164	Nagy James	Emory University Atlanta - USA nagy@mathcs.emory.edu
165	Nakatsukasa Yuji	University of Tokyo Tokyo - Japan nakatsukasa@mist.i.u-tokyo.ac.jp
167	Ng Esmond	Lawrence Berkeley National Laboratory Berkeley - U.S.A. EGNg@lbl.gov
169	Nichols Nancy	University of Reading Reading - UK n.k.nichols@reading.ac.uk

continued . . .

. . . continued

Page	Speakers	Misc
170	Noferini Vanni	The University of Manchester Manchester - United Kingdom vanni.noferini@manchester.ac.uk
171	Overton Michael	New York University New York - USA overton@cs.nyu.edu
173	Paige Christopher	McGill University Montreal - Canada paige@cs.mcgill.ca
175	Parlett Beresford	University of California, Berkeley Berkeley - USA parlett@math.berkeley.edu
176	Pearson John	University of Edinburgh Edinburgh - UK j.pearson@ed.ac.uk
178	Pestana Jennifer	University of Oxford Oxford - England pestana@maths.ox.ac.uk
180	Plešinger Martin	Technical University of Liberec Liberec 1 - Czech Republic martin.plesinger@tul.cz
182	Plestenjak Bor	University of Ljubljana Ljubljana - Slovenia bor.plestenjak@fmf.uni-lj.si
183	Poloni Federico	Università di Pisa Pisa - Italy fpoloni@di.unipi.it
186	Ramage Alison	University of Strathclyde Glasgow - Scotland A.Ramage@strath.ac.uk
187	Rees Tyrone	STFC Rutherford Appleton Laboratory Didcot - United Kingdom tyrone.rees@stfc.ac.uk
188	Reichel Lothar	Kent State University Kent - United States reichel@math.kent.edu
190	Renaut Rosemary	Arizona State University Tempe - United States renaut@asu.edu
192	Rozložník Miro	Czech Academy of Sciences Prague - Czech Republic miro@cs.cas.cz

continued . . .

. . . continued

Page	Speakers	Misc
194	Ruhe Axel	KTH, Royal Institute of Technology Stockholm - Sweden ruhe@kth.se
195	Rump Siegfried	Hamburg University of Technology Hamburg - Germany rump@tuhh.de
196	Sartenaer Annick	University of Namur Namur - Belgium annick.sartenaer@unamur.be
198	Saunders Michael	Stanford University Stanford - United States saunders@stanford.edu
199	Schröder Christian	Technical University Berlin Berlin - Germany schroed@math.tu-berlin.de
201	Scott Jennifer	STFC Rutherford Appleton Laboratory Didcot - UK jennifer.scott@stfc.ac.uk
203	Shao Meiyue	EPF Lausanne Lausanne - Switzerland meiyue.shao@gmail.com
205	Sifuentes Josef	Texas A&M University Bryan - USA. josefs@math.tamu.edu
207	Solomonik Edgar	University of California, Berkeley Berkeley - USA solomon@eecs.berkeley.edu
209	Soodhalter Kirk	Johannes Kepler University Linz - Austria kirk@math.soodhalter.com
211	Sorensen Danny	Rice University Houston - United States sorensen@rice.edu
212	Stewart G	University of Maryland College Park - USA stewart@cs.umd.edu
214	Stoll Martin	MPI Magdeburg Magdeburg - Germany stollm@mpi-magdeburg.mpg.de
216	Strakoš Zdeněk	Charles University in Prague Prague 8 - Czech Republic strakos@karlin.mff.cuni.cz

continued . . .

. . . continued

Page	Speakers	Misc
218	Stykel Tatjana	University of Augsburg Augsburg - Germany stykel@math.uni-augsburg.de
220	Szyld Daniel	Temple University Philadelphia - USA szyld@temple.edu
222	Taslaman Leo	The University of Manchester Manchester - United Kingdom leotaslaman@gmail.com
224	Tichý Petr	Czech Academy of Sciences Prague - Czech Republic tichy@cs.cas.cz
226	Tisseur Francoise	The University of Manchester Manchester - United Kingdom francoise.tisseur@manchester.ac.uk
228	Titley-Peloquin David	CERFACS Toulouse - France titleypelo@cerfacs.fr
230	Tuma Miroslav	Academy of Sciences of the Czech Republic Praha - Czech Republic tuma@cs.cas.cz
232	André Uschmajew	EPF Lausanne Lausanne - Switzerland andre.uschmajew@epfl.ch
234	Van Barel Marc	KU Leuven Leuven (Heverlee) - Belgium marc.vanbarel@cs.kuleuven.be
235	Van Beeumen Roel	KU Leuven Heverlee - Belgium Roel.VanBeeumen@cs.kuleuven.be
237	Vandebril Raphael	KU Leuven Leuven - Belgium Raf.Vandebril@cs.kuleuven.be
238	Vandereycken Bart	Princeton University PRINCETON - United States bartv@math.princeton.edu
240	Van Dooren Paul	Université catholique de Louvain Louvain-la-Neuve - Belgium paul.vandooren@uclouvain.be
241	Van Huffel Sabine	KU Leuven Leuven - Belgium sabine.vanhuffel@esat.kuleuven.be

continued . . .

. . . continued

Page	Speakers	Misc
243	Van Loan Charles	Cornell University Ithaca - USA cv@cs.cornell.edu
244	Vassilevski Panayot	Lawrence Livermore National Laboratory Livermore - USA panayot@llnl.gov
246	Vavasis Stephen	University of Waterloo Waterloo - Canada vavasis@uwaterloo.ca
248	Voigt Matthias	Max Planck Institute Magdeburg - Germany voigtm@mpi-magdeburg.mpg.de
250	Wathen Andy	Oxford University Oxford - UK wathen@maths.ox.ac.uk
251	Watkins David	Washington State University Pullman - USA watkins@math.wsu.edu
252	Wei Yimin	Fudan University Shanghai - P.R. China yimin.wei@gmail.com
253	Xu Hongguo	University of Kansas Lawrence - USA xu@math.ku.edu
254	Ziętak Krystyna	Wroclaw School of Information Technology Wroclaw - Poland k.zietak@gmail.com