Model Correction using a Nuclear Norm Constraint

Ning Hao, Lior Horesh, Misha Kilmer

Ragon Institute
IBM Thomas J. Watson Research Center
Tufts University
Thanks: NSF:CIF:SMALL 1319653, NSF-DMS 0914957

Householder Symposium XIX, June 2014

Motivation

We create mathematical models to simulate the operation of a real world or system.

Not surprisingly, the models will contain some error:

- simplifications for computational tractability
- linearization, numerical error
- boundary conditions or geometry

Motivation

We create mathematical models to simulate the operation of a real world or system.

Not surprisingly, the models will contain some error:

- simplifications for computational tractability
- linearization, numerical error
- boundary conditions or geometry

"All models are wrong . . . but some are useful "
George E. P. Box, 1987

Motivation

We create mathematical models to simulate the operation of a real world or system.

Not surprisingly, the models will contain some error:

- simplifications for computational tractability
- linearization, numerical error
- boundary conditions or geometry

"All models are wrong . . . but some are useful "
George E. P. Box, 1987

Thinking outside the Box, some models can be made less wrong / more useful. This is the subject of today's talk.

Problem Definition

Let $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^m$ be "true" operator taking input $x \in \mathbb{R}^n$ to observable space. Let $d \in \mathbb{R}^m$ be an observation obtained as

$$d = \mathcal{F}(x) + \epsilon$$

where ϵ represents measurement noise.

Goal: Formulate and solve a model correction optimization problem to recover an $\mathcal{F}$ that is only partially specified.
In this talk, assume A is known and

$$\mathcal{F}(x) = A(x) + B(x).$$

Model correction problem is to recover B. Need additional constraints on B to obtain a well-posed optimization problem.

Motivating Test Problem

Discrete ill-posed problem (e.g. deblurring, Xray-CT)

$$Ax = d + \epsilon.$$

If operator not known exactly model should be replaced by

$$(A + B)x = d + \epsilon.$$

While B is unknown, and in general x is unknown, we assume we can sample input space and have access to corresponding output realizations for each of the samples.

Sample Average Approximation

SAA[1] : Expectation wrt the input space & the measurement noise approximated by a sample ave. est. of a random sample.

$$\hat{B} = \underset{B}{\operatorname{argmin}} \frac{1}{n_x n_\epsilon} \sum_{j=1,j=1}^{n_x,n_\epsilon} \mathcal{D}\left((A(x_i) + B(x_i)) - d_{ij}\right)$$

subject to constraint on B, n_x is the number of input draws, n_ϵ is number of data realizations for each input realization.

Problem ($A(x) = Ax, B(x) = Bx, \mathcal{D}(y) = \|y\|_2^2$)

$$\hat{B} = \arg\min_{B} \frac{1}{n_x n_\epsilon} \sum_{i=1,j=1}^{n_x,n_\epsilon} \|(A+B)x_i - d_{i,j}\|_2^2,$$

subject to a structural constraint on B .

[1] Shapiro, Dentcheva, Ruszczynski, editors. "Lecture Notes on Stochastic Programming: Modeling and Theory." SIAM, 2009.

Restrictions on Operator Structure

Many constraints on B possible. Choose rank-based:

Problem

$$\hat{B} = \underset{B}{\operatorname{argmin}} \|(A+B)X - D\|_F^2$$

$$\textit{s.t.} \quad \textit{rank}(B) \leq \frac{\delta}{2}$$

Relax the constraint:

Problem

$$\hat{B} = \underset{B}{\operatorname{argmin}} ||(A+B)X - D||_F^2$$

$$\textit{s.t.} \quad \|B\|_* \leq \frac{\delta}{2}$$

Optimization Problem with Nuclear Norm Regularization

Several options for solving the optimization problem[1]

- Interior Point Methods
- Alternating Direction Method of Multipliers (ADMM)
- Projected Sub-Gradient Method
- More recent work by Jaggi and Sulovsky: recast the optimization problem over positive semidefinite matrices with unit trace and then apply Hazan's algorithm.

[1] Hao, Horesh, & K., "Nuclear norm optimization and its application to observation model specification," in Compressed Sensing and Sparse Filtering, 2014

Hazan's Algorithm

Hazan's Algorithm deals with problems of the form:

$$\min_{Z \in \mathcal{S}} f(Z)$$

where f is convex and $\mathcal{S}$ is the set of all symmetric positive semidefinite $d \times d$ matrices with unit trace.

Each iteration involves the calculation of a single approximate eigenvector of a matrix of size of $d \times d$.

Sparse approximation solutions to semidefinite programs by Hazan,E. 2008

Algorithm

From Jaggi and Sulovsky "A Simple Algorithm for Nuclear Norm Regularized Problem" 2010, for any non-zero matrix $B \in \mathbb{R}^{n\times m}$ and $\delta \in \mathbb{R}$:

$$\|B\|_* \leq \frac{\delta}{2} \iff \exists \text{ symmetric } M, N \text{s.t.}$$

$$\begin{pmatrix} M & B \\ B^T & N \end{pmatrix} \succeq 0$$

and

$$\mathrm{Tr}(M) + \mathrm{Tr}(N) = \delta.$$

Steps

Let $Z = \begin{pmatrix} M & B \\ B^T & N \end{pmatrix}$.

- Define $\hat{f}(Z) = f(B) = \|(A+B)X - D\|_F^2$,
- Want to solve

$$\min_Z \hat{f}(Z)$$

$$\text{s.t. } Z \in S^{(m+n)\times(m+n)}, Z \succeq 0, \mathsf{Tr}(Z) = \delta$$

- Scale all matrix entries by $\frac{1}{\delta}$
- Apply Hazan's algorithm, then unscale

Algorithm

Input: f, $v_0 \in \mathbb{R}^{(m+n)\times 1}$ with $||v_0|| = 1$; scaled matrices
Initialize: $Z_1 = v_0 v_0^T$
for $k = 1$ until convergence do
    Extract $B_k = Z_k(1:m, m{+}1{:}m{+}n)$
    Compute $\nabla f_k := ((A + B_k)X - D)X^T$
    We have $\nabla \hat{f}_k := \begin{pmatrix} 0 & \nabla f_k \\ \nabla f_k^T & 0 \end{pmatrix}$
    Compute $v_k :=$ eigs$(-\nabla \hat{f}_k, 1,' LA')$
    Line search for step length a_k
    Update $Z_{k+1} = Z_k + a_k(v_k v_k^T - Z_k)$
end for
Return $\hat{B} = \delta \cdot Z(1{:}m, m{+}1{:}m{+}n)$

Numerical Examples

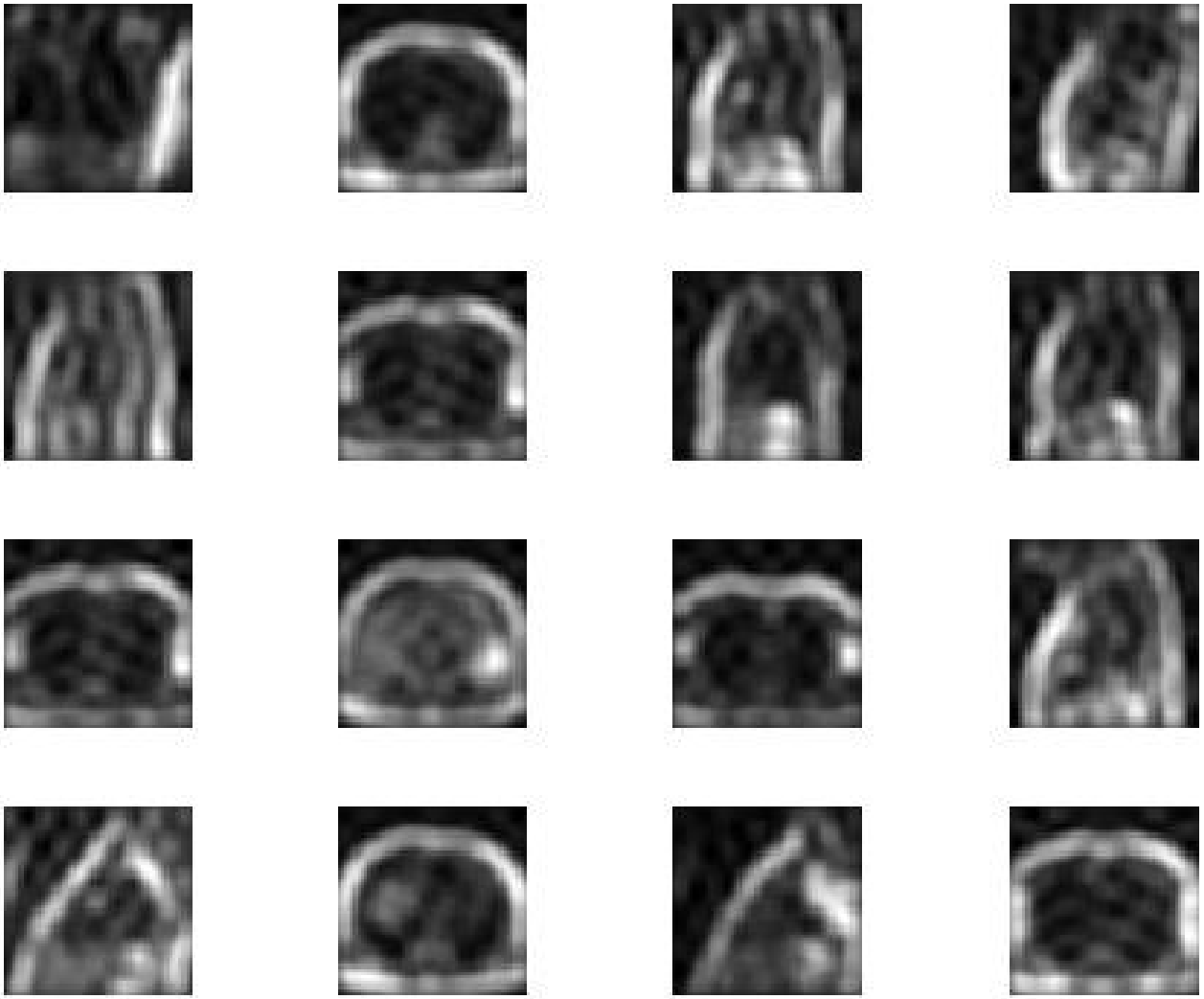
- Mimic the semi-blind deconvolution problem, only an approximation to blurring operator is known a-priori.
- 100 MR images from the Auckland MRI Research group database.
- Pre-process the images by cropping the watermark and resizing the cropped images to the size of 55×55.
- Randomly choose 80 images (sampled integers from 1:100 uniformly) to serve as the training set and 20 of remainder as the test set.

http://atlas.scmr.org/download.html

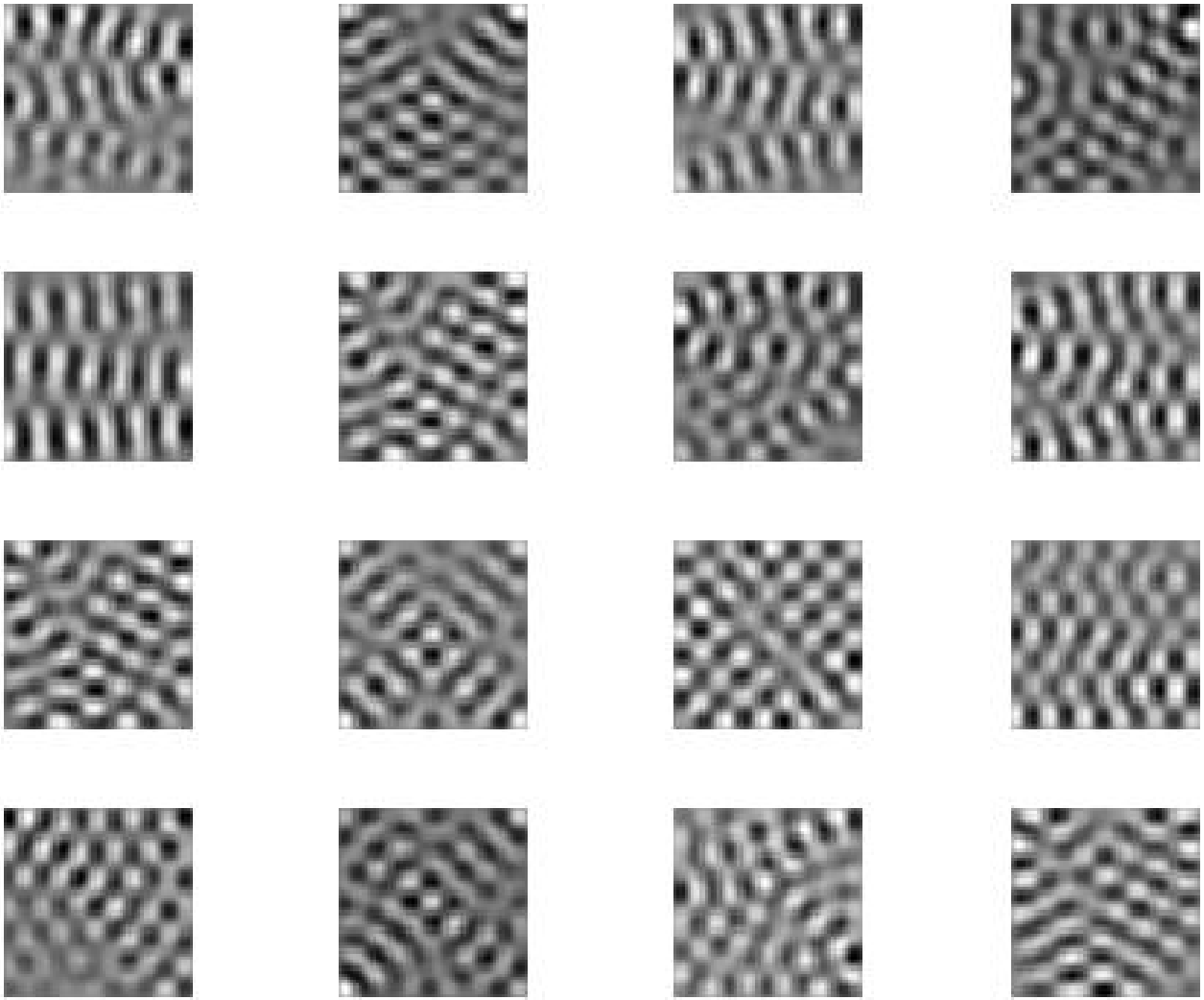
Example 1

- Let $T(b_w, \sigma)$ be symmetric, doubly block Toeplitz matrix T that models blurring of an image by a Gaussian point spread function; b_w, σ control (block) bandwidth and blur.
- $A = T(4, 4)$
- B defined from singular triples of $T(3, 2)$ numbered 110 to 120.
- Compute $D = (A + B)X + N$ where N's columns have zero mean, white Gaussian noise such that the noise to signal ratio for all measured data is .5 percent.
- Run 120 iterations. Compare the best TSVD regularized solutions using the SVDs of $A, A + B, A + \hat{B}$.

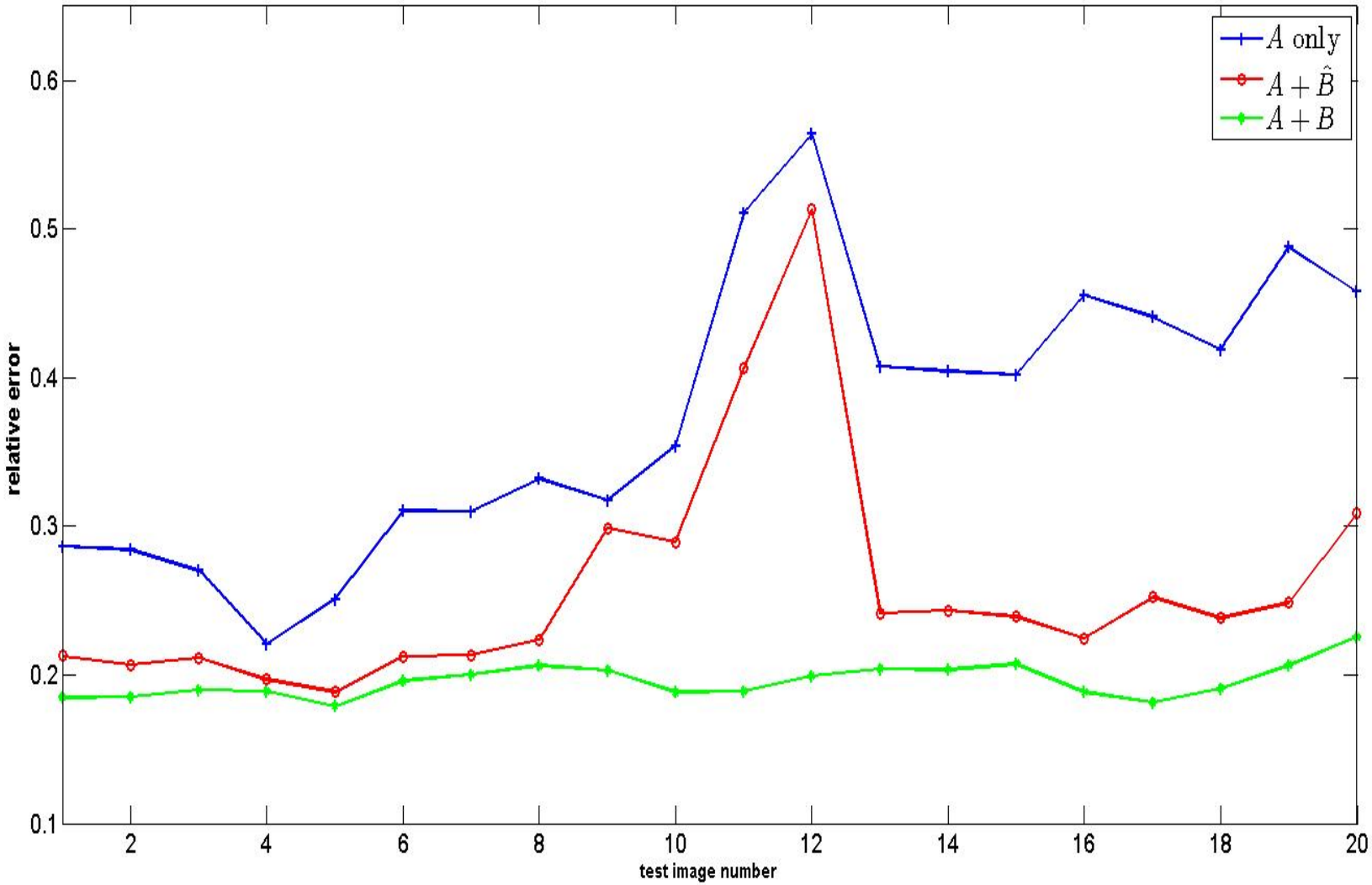
True Blurred Training Images



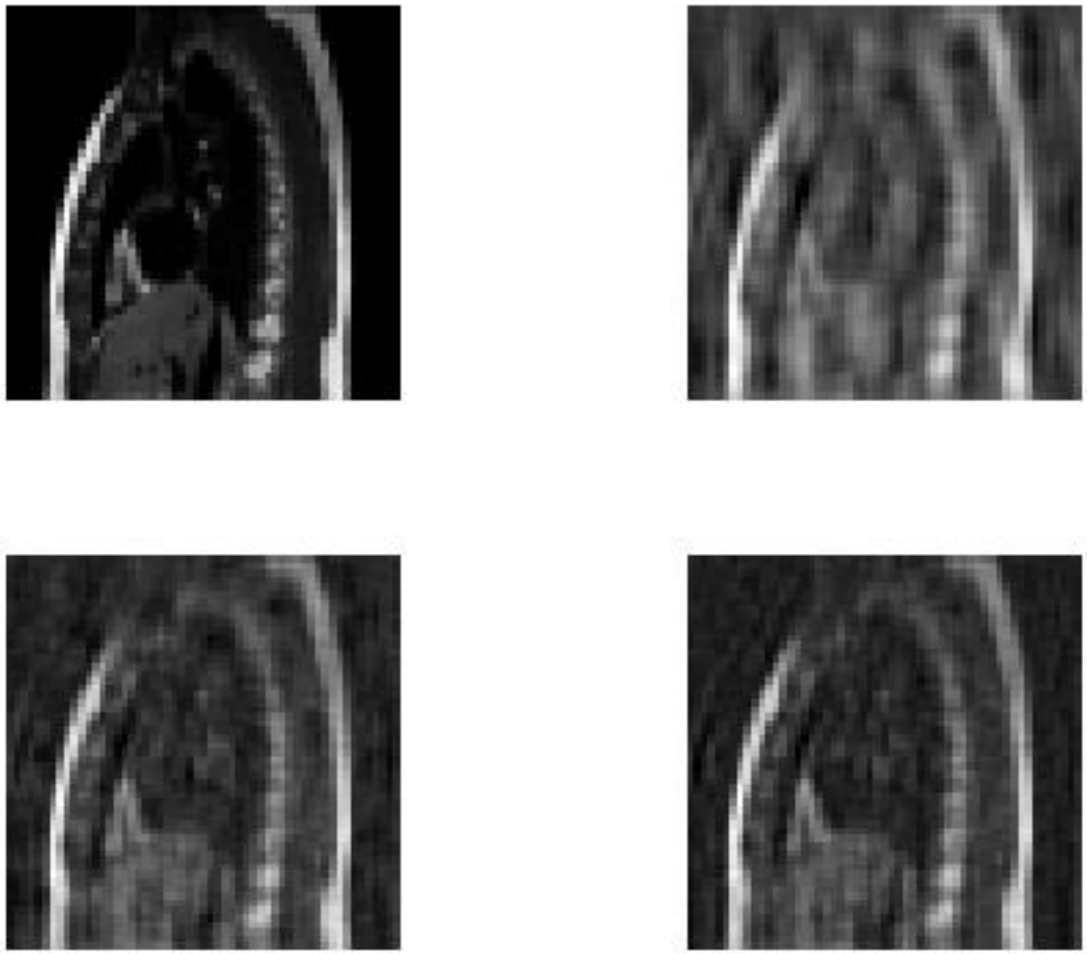
AX **vs** $(A+B)X$



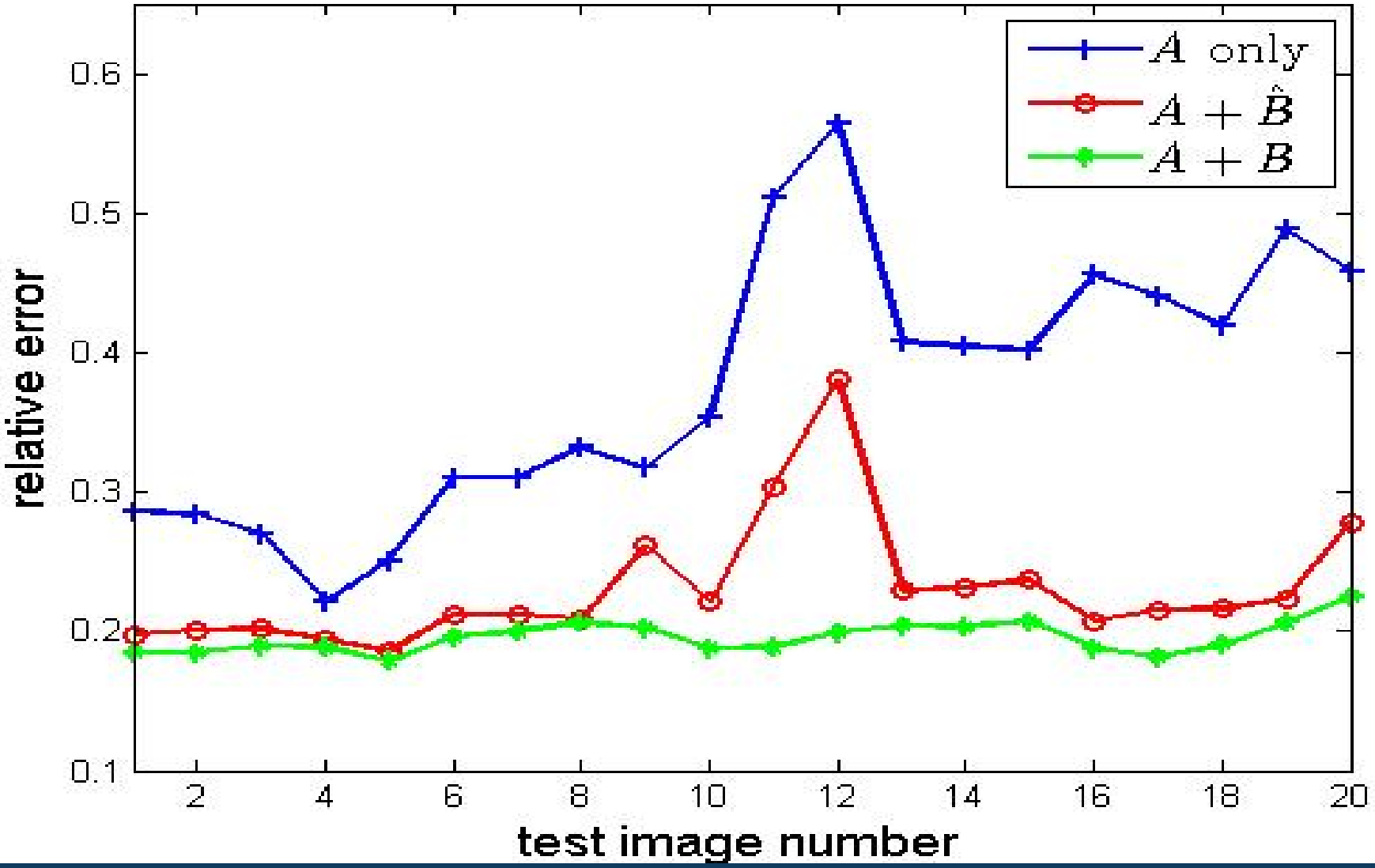
Rel. Errors, TSVD recons, 20 from test set



Example 1, Test Slice 16



Rel. Errors after 400 iterations



Enforcing Different Structure

What if $B(x) \neq Bx$, but rather $B(x) = K[B]x$?

Example [Kamm & Nagy, 1998] Let v be of length n, n odd, toep(v) creates an $n \times n$ banded Toeplitz matrix with $v(\frac{n+1}{2})$ as the diagonal entry:

$$\text{toep}\left(\begin{bmatrix} a \\ b \\ c \end{bmatrix}\right) = \begin{bmatrix} b & a & 0 \\ c & b & a \\ 0 & c & b \end{bmatrix}$$

Additional Structure from B

Given $n \times n$ B, make a BTTB matrix from blocks toep$B(:, i)$ in a similar manner.

For B with 3 columns:

$$K[B] = \begin{bmatrix} \text{toep}(B(:,2)) & \text{toep}(B(:,1)) & 0 \\ \text{toep}(B(:,3)) & \text{toep}(B(:,2)) & \text{toep}(B(:,1)) \\ 0 & \text{toep}(B(:,3)) & \text{toep}(B(:,2)) \end{bmatrix}.$$

Convexity, Constraint Revisited

It is possible to show that
$K[B]x = K[\text{reshape}(x, n, n)]\text{vec}(B)$.

$$\text{So,} \qquad \sum \|Ax_{ij} + B(x_{ij}) - d_{ij}\|_F^2$$

still convex in entries of B

Significance of the (ideal) rank-based constraint on B:

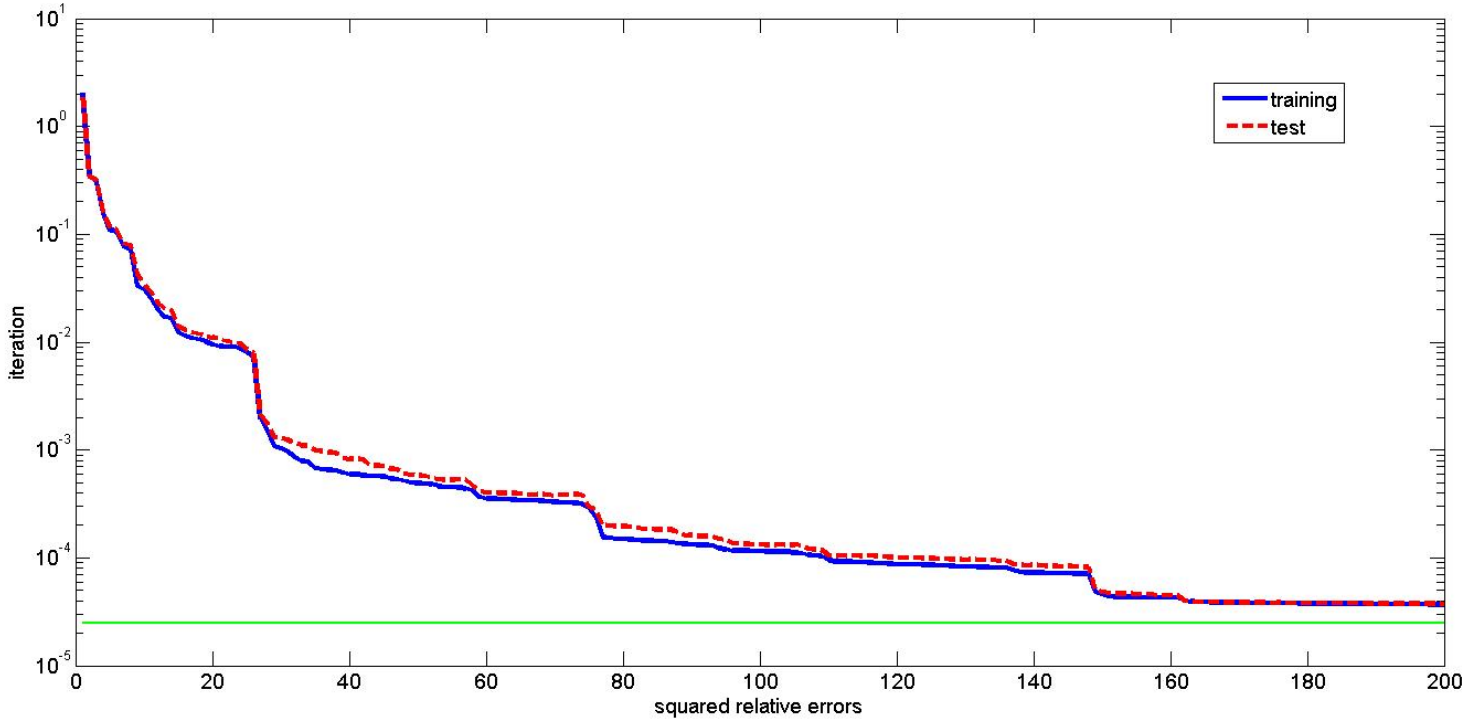
$$B = \sum_{i=1}^{k} \sigma_i u_i v_i^T \Rightarrow K[B] = \sum_{i=1}^{k} \text{toep}(\sqrt{\sigma_i} v_i) \otimes \text{toep}(\sqrt{\sigma_i} u_i),$$

Algorithm now cheap - matrix of partials is now $n \times n$, matvecs with FFTs.

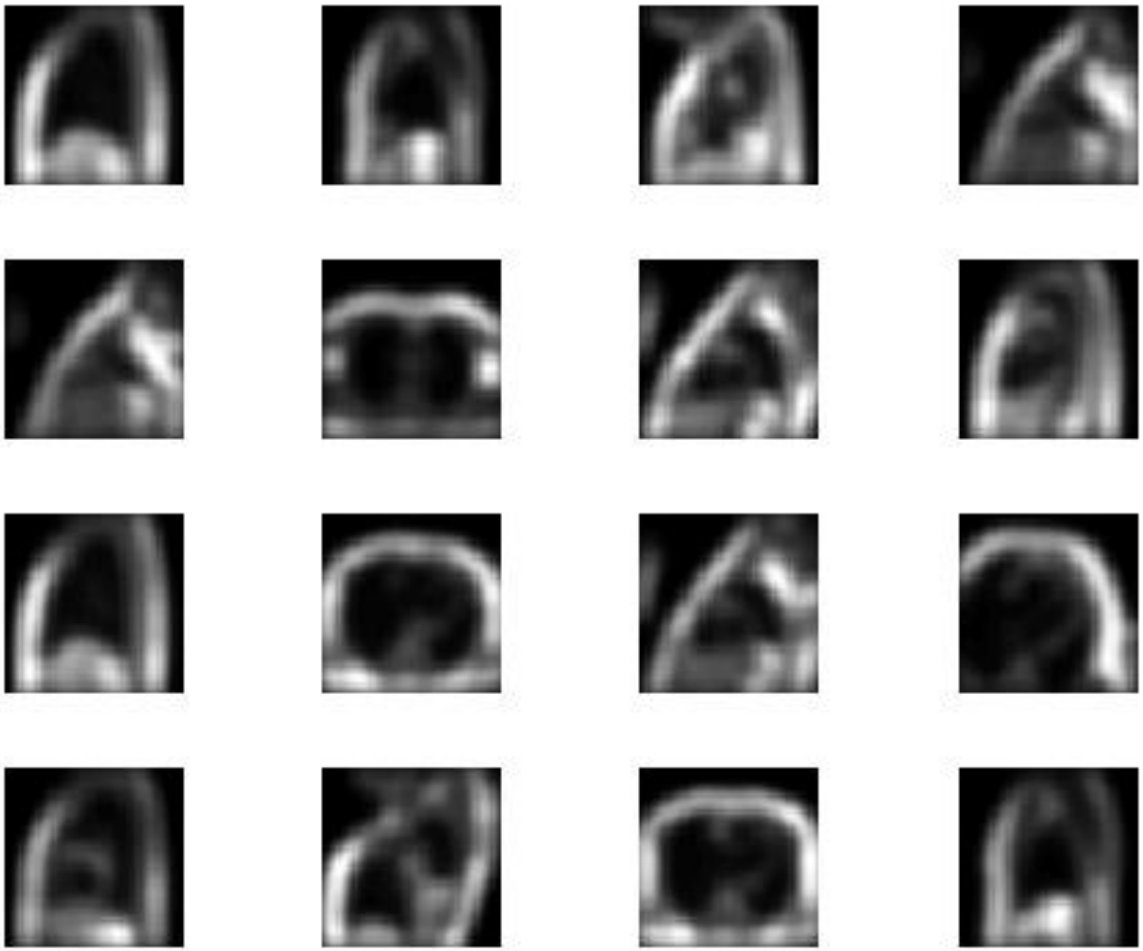
Example 2

- $A = T(10, 4)$
- $B = T(7, 3.1) + T(4, 2.2) + T(5, 3.5)$
- Same testing setup as before (random selection of 80 images) and testing on the remainder, Gaussian $\frac{1}{2}$ percent noise added (once per test image).

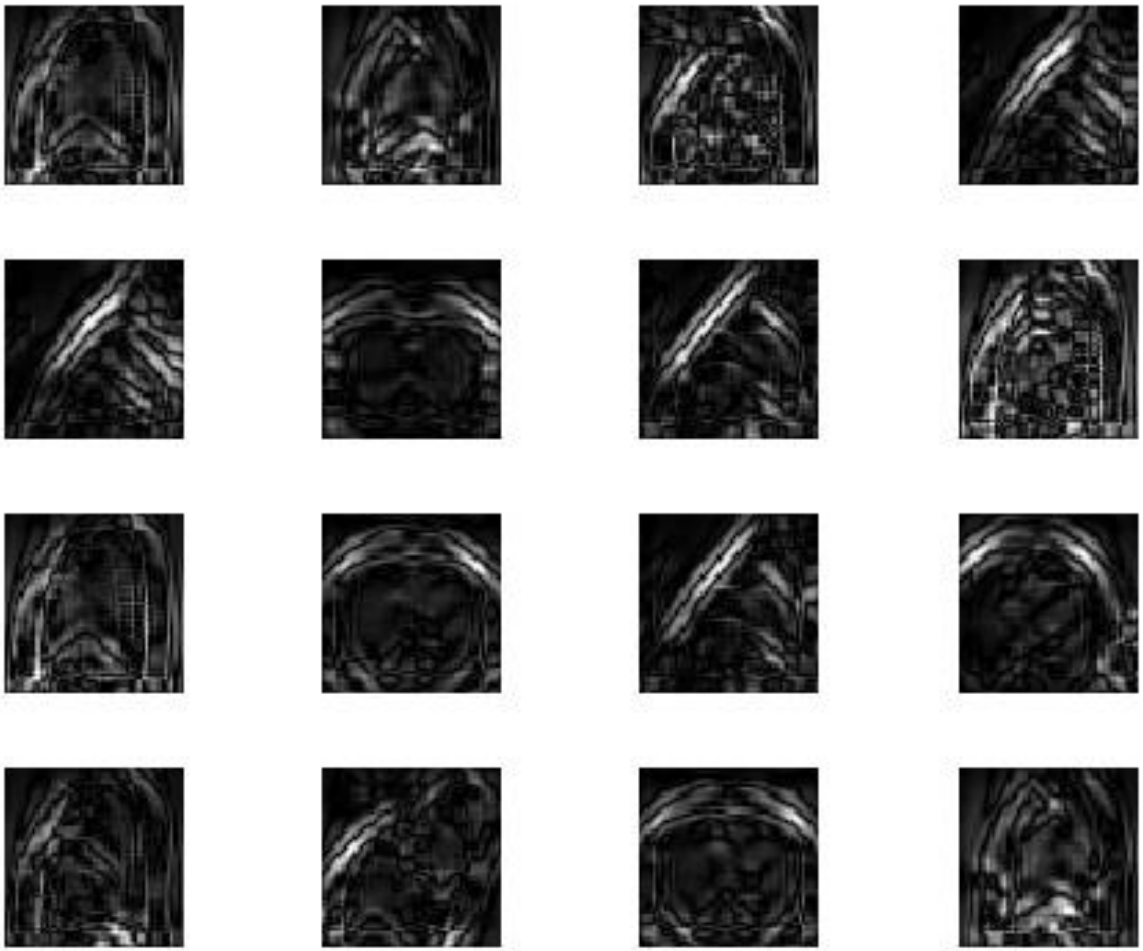
Relative Squared Convergence Error



Difference in Data Space: $|BX_{train}|$

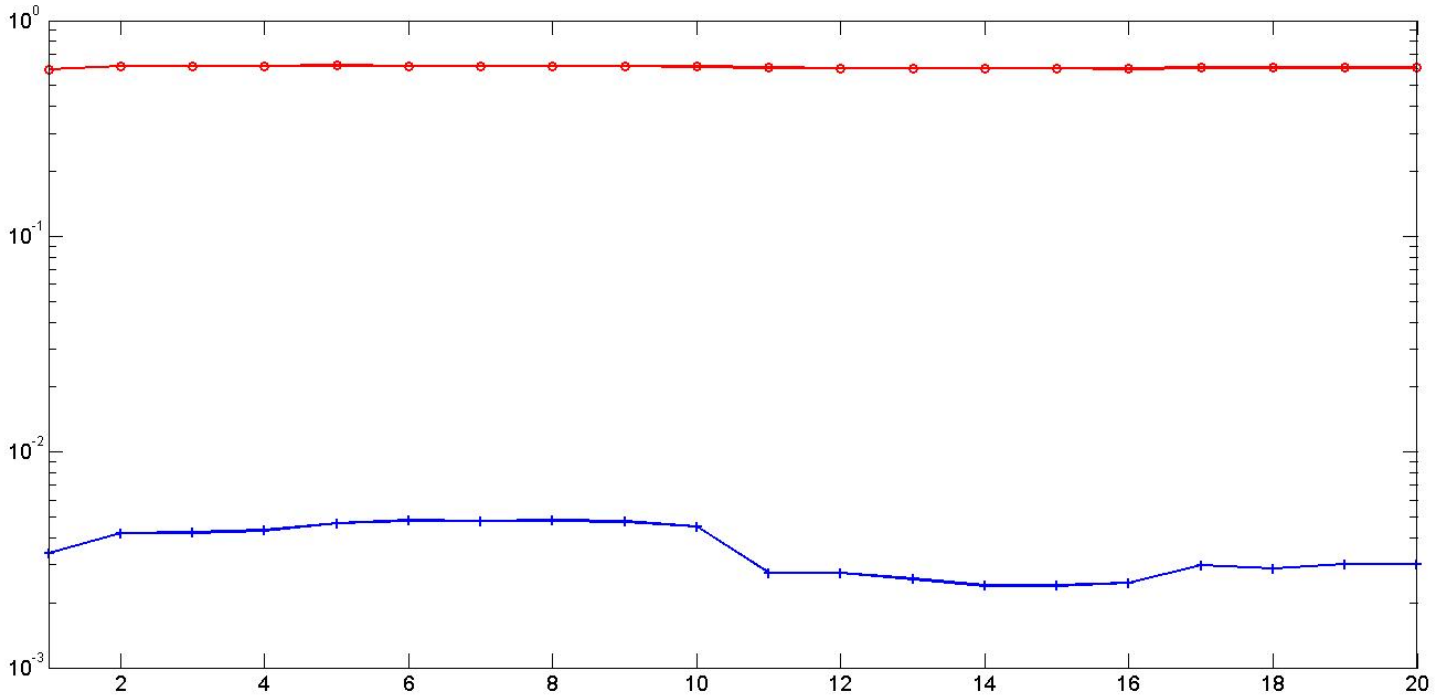


Difference in Data Space: $|(B - \hat{B})X_{train}|$



Errors on 20 Test Data

$$\frac{\|AX-(A+B)X\|_F}{\|(A+B)X\|_F} \text{ vs. } \frac{\|(\hat{B}-B)X\|_F}{\|(A+B)X\|_F}$$



One Extension to Tensors

B is order > 2 tensor from which we define $K[B](x)$.

$B \in \mathbb{R}^{n \times n \times n}$, $B = \sum_{i=1}^{r} u_i \circ v_i \circ w_i \Rightarrow K[B]$ to a sum of 3-way Kronecker products of structured matrices

- Nagy & K., "Kronecker Product Approximation for Three-Dimensional Imaging Applications, IEEE TIP 2006.
- Rezghi & Elden, "A Kronecker Product Approximation of the Blurring Operator in the Three Dimensional Image Restoration Problem," SIMAX to appear.

But what do we mean by $\|B\|_*$?

One Extension to Tensors

B is order > 2 tensor from which we define $K[B](x)$.

$B \in \mathbb{R}^{n \times n \times n}$, $B = \sum_{i=1}^{r} u_i \circ v_i \circ w_i \Rightarrow K[B]$ to a sum of 3-way Kronecker products of structured matrices

- Nagy & K., "Kronecker Product Approximation for Three-Dimensional Imaging Applications, IEEE TIP 2006.
- Rezghi & Elden, "A Kronecker Product Approximation of the Blurring Operator in the Three Dimensional Image Restoration Problem," SIMAX to appear.

But what do we mean by $\|B\|_*$?

More thinking outside the Box....

One Extension to Tensors

B is order > 2 tensor from which we define $K[B](x)$.

$B \in \mathbb{R}^{n \times n \times n}$, $B = \sum_{i=1}^{r} u_i \circ v_i \circ w_i \Rightarrow K[B]$ to a sum of 3-way Kronecker products of structured matrices

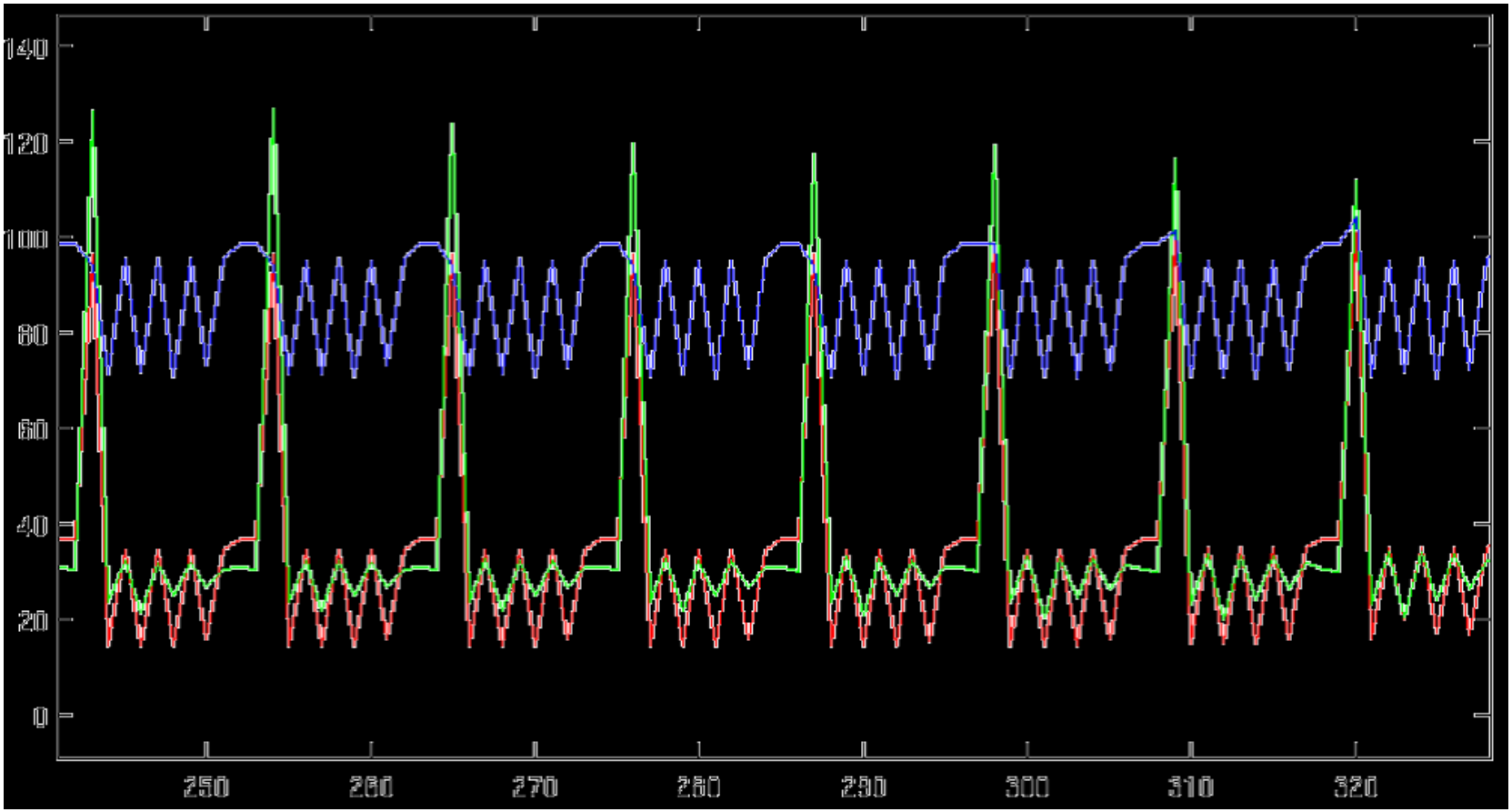
- Nagy & K., "Kronecker Product Approximation for Three-Dimensional Imaging Applications, IEEE TIP 2006.
- Rezghi & Elden, "A Kronecker Product Approximation of the Blurring Operator in the Three Dimensional Image Restoration Problem," SIMAX to appear.

But what do we mean by $\|B\|_*$?

Use $\|B\|_{\text{TNN}}$ definition [Hao, 2014] based on tSVD [K. & Martin, 2011].

Current & Future Work

Other examples (PDEs) (thanks: Raya Horesh, IBM Watson)



Current & Future Work

- Tensors
- Other examples (PDES)
- Other choices for $\mathcal{D}$
- Investigate other possible constraints on the operator (nonlinear correction).