Partial Least Squares and Cox model with application to gene expression

Sophie Lambert-Lacroix,
UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG
UMR 5525, Grenoble, F-38041, France
and Frédérique Letué,
LJK, Université de Grenoble et CNRS, UMR 5224
51, rue des Mathématiques, B.P. 53, 38041 Grenoble cedex 9, FRANCE

July 26, 2011

Abstract

One important aspect of data-mining of microarray data is to discover the molecular variation among cancers. In microarray studies, the number n of samples is relatively small compared to the number p of genes per sample (usually in thousands). That is a considerable challenge in the context of survival prediction. This naturally calls for the use of a dimension reduction procedure together with the prediction one. In this paper, the question of survival prediction in such a high dimensional setting is addressed. We propose a new method combining Partial Least Squares (PLS) and Ridge penalized Cox regression. We review the existing methods based on PLS and (or) penalized likelihood techniques, outline their interest in some cases and theoretically explain their sometimes poor behavior. Our procedure is compared with these other methods. The performance of the resulting procedures is illustrated on two real data sets.

Availability: R codes are available upon request.

Keywords: Cox model, dimension reduction procedure, microarray data, partial least squares, survival prediction.

1 Introduction

Microarray technology generates a vast amount of data by measuring, through the hybridization process, the levels of virtually all the genes expressed in a biological sample. One can expect that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine.

In survival analysis, survival times as time to cancer recurrence or death due to cancer are studied. The main goal is to predict the time to event (survival time) using the gene expressions as covariables. One first difficulty to study time to event outcome results from right censoring during patient follow-up since some patients may still be event-free. Such observations are said to be right-censored; for these patients, we only know that the time to event is greater than the time of last follow-up. So standard regression techniques cannot be applied since the event is not observed for all samples. We consider here only the methods that use the proportional hazard model introduced by Cox [4] to link the survival time to gene expressions. However, the Cox proportional hazard model is usually applied in situation where the number of samples, n, greatly exceeds the number of covariates, p. In microarray studies, n is relatively small compared to the number p of genes, usually in thousands ($p >> n$). In addition, gene expressions data are often highly correlated. That is a considerable challenge in the context of survival prediction.

Similar data structures have been encountered in the field of chemometrics. The method of Partial Least Squares (PLS, [25, 15, 9]) has been found to be a useful dimension reduction technique as well as Principal Component Regression (PCR, [14]) (see [7] for a statistical view of PLS and PCR). In the context of microarrays, the purpose of PCR or PLS is to produce orthogonal tumor descriptors that reduce the dimension to only few gene components (super-genes). But the dimension reduction for PCR is achieved without regard to the response variable and may be inefficient.

Nguyen and Rocke [16] propose to apply PLS directly to the survival data and use the resulted PLS components in the Cox regression model to predict survival time. However, their procedure does not handle the censoring aspect properly. There exist some other extensions that try to take into account this fact. In [18], the authors reformulate the Cox model as a Poisson model for the censored indicator variable. Then they use the PLS algorithm developed by [13] in the generalized linear model. The reformulation as a Poisson regression increases the dimension of the problem and in high dimensional data this algorithm may fail to converge (see [12]). We do not also consider this approach.

In Bastien [2], the standard PLS method is modified by replacing the linear

regression step by Cox regression to determinate the PLS components. The first component is obtained by a weighted sum of centered expression values. Next, these expression values are regressed against the PLS components and the residuals are used for building the next component (in a similar fashion). The procedure is repeated until all the components are constructed. Li and Gui [12] propose a similar approach. The difference stands in the choice of the weights.

To deal with the high-dimensional problem, another approach consists in penalizing the Cox's Partial log-likelihood. In [8, 22], the authors propose to use the Ridge penality in order to both stabilize the statistical problem and remove numerical degeneracy due to multicollinearity: a so-called Ridge penalty is subtracted from the Cox's Partial log-likelihood. Note that this method is not a dimension reduction technique. Indeed all explanatory variables are allowed into the regression model. Indeed all the genes contribute, which can inhibit and degrade the performances of the prediction rules.

In this paper, we compare several dimension reduction and/or regularization methods for the Cox model. These methods are: Ridge Cox, Cox PCR, both PLS methods for the Cox model proposed by Li and Gui [12] and Bastien [2], Cox Lasso [21]. In addition, we extend the PLS method to Cox regression in a similar way to the extension to generalized linear model proposed by [6]. To do that, the idea is to substitute the survival time in the input of PLS by a continuous-valued pseudo-response variable whose expected value has a linear relationship with the covariates. The limiting pseudo-response variable in the Iteratively Reweighted Least Squares (IRLS) algorithm used to compute the maximum Cox's Partial log-likelihood seems to be a good candidate. Unfortunately, in the present situation "small n, large p", IRLS no longer works since the limiting pseudo-response variable is, in norm, infinite. The idea developed here is to penalize with a Ridge penalty the Cox's partial log-likelihood criterion in order to constrain the pseudo-response variable to be finite. Our procedure combines a Ridge penalty step and a PLS step and the output of the dimension reduction step is incorporated in the Cox regression step.

This paper is organized as follows. The Methods section is the methodological part of this paper. It contains a description of the Cox regression. We then recall the Ridge Cox's partial log-likelihood method and derive a weighted PLS algorithm in order to address the dimension reduction in heteroscedastic models. We then introduce an extension of PLS to survival time data based on the Ridge penalty. We focus on the method of Li and Gui [12] and that of Bastien [2]. We show that both methods can be described by the same algorithm. The difference between both approaches lies in the choice of some weights. Applications to survival prediction using gene expression data are presented in the Numerical results section.

2 Methods

2.1 Cox model for survival data

We consider the usual survival data setup. Let $(T_1, \ldots, T_n)$ be independent survival times, and $(U_1, \ldots, U_n)$ be censoring times. We observe the p-dimensional vectors of covariates $\underline{X}_1, \ldots, \underline{X}_n$. The right censored survival time is given by $\tilde{T}_i = min(T_i, U_i)$. We denote by $\Delta_i = \mathbb{1}(T_i \leq U_i)$ the indicator of event and τ the study cutoff time. In the survival data setup, we observe n i.i.d. copies $(\tilde{T}_i, \Delta_i, \underline{X}_i)$ of $(\tilde{T}, \Delta, \underline{X})$, $i = 1, \ldots, n$. We denote by $(N_i, i = 1, \ldots, n)$ the corresponding counting processes: $N_i(t) = \mathbb{1}(T_i \leq t, \Delta_i = 1)$. In the microarray gene context, $\underline{X}_1, \ldots, \underline{X}_n$ are the expression levels of p genes and $(T_1, \ldots, T_n)$ are the survival times such as time to cancer recurrence or death due to cancer.

We consider the following Cox regression model for the hazard function,

$$\alpha_{\underline{X}_i}(t) = \alpha_0(t) \exp(\underline{X}_i^T \underline{\beta}),$$

where $\underline{\beta}$ is the p-dimensional vector of parameters and α_0 is a baseline hazard function. Let us note that this model does not contain any intercept term since the multiplicative term $\lambda_0(t)$ may contain this term.

To estimate the parameters vector $\underline{\beta}$, one usually maximizes the Cox's Partial Log-likelihood (PL) given by

$$l_n(\underline{\beta}) = \sum_{i=1}^{n} \int_0^\tau \log \frac{e^{\underline{X}_i^T \underline{\beta}}}{S_n(\underline{\beta}, s)} dN_i(s)$$

where $S_n(\underline{\beta}, s) = \sum_{j=1}^{n} \mathbb{1}(T_j \geq s) \exp(\underline{X}_j^T \underline{\beta})$.

Note that the likelihood does not change when a constant is added to the covariates. Therefore, there is no need to center the covariates in this context. In the sequel, we need some additional notations. Expression levels of the p genes for the n microarray samples are collected in a $n \times p$ *data matrix* $\mathbf{X} = (x_{ij})$, $1 \leq i \leq n, 1 \leq j \leq p$. The entry x_{ij} is the expression level of the variable "gene" j in the microarray sample i, and the i-th row $\mathbf{X}_{i,\cdot}$ is the vector of a gene expression profile for sample i. The vector $\underline{\tilde{T}}$ is the n-dimensional vector of the right censored survival time and $\underline{\tilde{\Delta}}$ the corresponding vector of the indicator of event.

2.2 Maximum PL estimate and Iteratively Reweighted Least Squares (IRLS)

We say that the PL estimate exists if there exists $\underline{\beta} \in \mathbb{R}^p$ of finite norm which is a maximizer of the concave partial log-likelihood ℓ_n. Hence, such an estimate is a solution to the normal equation $\mathbf{X}^T \underline{\zeta}(\underline{\beta}) = 0$, where $\underline{\zeta}(\underline{\beta})$ is defined by

$$\underline{\zeta}_k(\underline{\beta}) = \int_0^\tau \sum_{i=1}^n (\delta_{i,k} - w_k(\underline{\beta}, s)) dN_i(s), \tag{1}$$

with $\delta_{i,k}$ being the Kronecker symbol and the $w_k(\underline{\beta}, s)$ being weights equal to

$$\frac{\exp(\underline{X}_k^T \underline{\beta}) \mathbb{1}(T_k \geq s)}{S_n(\underline{\beta}, s)}.$$

For a reference, see for instance [17]. Here, we just add the fact that the data are only observed in the interval $[0, \tau]$.

If $\mathbf{X}$ is full column-rank, the solution, when exists, is unique. In such a case, the estimate is usually computed as the limit of a converging Newton-Raphson sequence; this algorithm is known as the Iteratively Reweighted Least Squares (IRLS, see [10]). Let $\mathbf{W}(\underline{\beta})$ be the $n \times n$ matrix with entries given by

$$\mathbf{W}_{l,l'} = \int_0^\tau \sum_{i=1}^n w_l(\underline{\beta}, s)(\delta_{l,l'} - w_{l'}(\underline{\beta}, s)) d\bar{N}_i(s). \tag{2}$$

Each iteration divides into two steps,

$$\underline{z}^{(t)} = \mathbf{X}\underline{\beta}^{(t)} + \left[\mathbf{W}^{(t)}\right]^{-1} \underline{\zeta}, \tag{3}$$

$$\underline{\beta}^{(t+1)} = \left(\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \underline{z}^{(t)}, \tag{4}$$

where $\mathbf{W}^{(t)}$ is a shorthand notation for $\mathbf{W}(\beta^{(t)})$. The algorithm IRLS can thus be considered as an iterative weighted least square regression of a $\mathbb{R}^n$-valued pseudo-variable $z^{(t)}$ onto the columns of $\mathbf{X}$.

When $\mathbf{X}$ is not full column-rank, the parameter is not identifiable and the PL estimate is not unique when exists; applying the above iterations (3-4) by replacing the inverse matrix (4) with the Moore-Penrose pseudo-inverse, yields the parameter estimate which is of minimal norm among all the solutions. In practice, in the present statistical framework $n << p$, $n = \text{rank}(\mathbf{X})$ and the minimal norm solution verifies for all $1 \leq i \leq n$, $\zeta_i(\underline{\beta}) = 0$; it is thus of infinite norm and the PL estimate do not exist. This calls for regularization methods such as Ridge penalty.

2.3 Ridge penalty and RIRLS

There exist several studies that propose to use the Cox model with quadratic penalty to predict survival time based on gene expression data (see [8, 22]). The Ridge estimator $\hat{\underline{\beta}}^R$ is defined as the (unique) maximizer of the penalized likelihood

$$l_n^*(\underline{\beta}) = l_n(\underline{\beta}) - 0.5\lambda \underline{\beta}^T \underline{\beta},$$

where $\lambda > 0$ is the shrinkage parameter. The estimator $\hat{\underline{\beta}}^R$ always exists, is unique and is computed as the limit of a Newton-Raphson sequence. We denote by RIRLS $(\tilde{\underline{T}}, \tilde{\underline{\Delta}}, \tau, \mathbf{X}, \lambda)$ (shorthand notation for Ridge-IRLS) this algorithm. It consists in replacing in IRLS, the weighted regression (4) by a weighted Ridge regression

$$\underline{\beta}^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \underline{z}^{(t)},$$

where $\underline{z}^{(t)}$ is built as in (3) and I_p denotes the identity matrix of size $p \times p$. The parameter λ controls the amount of shrinkage in the data and can be chosen by cross-validation.

2.4 Weighted Partial Least Squares (WPLS)

Partial Least Squares (PLS) is both a tool for linear regression and a tool for dimension reduction [25, 15, 9]. Let $\underline{\mathbf{y}} \in \mathbb{R}^n$ be a response vector, $\mathbf{X}$ be a $n \times p$ data matrix and $\mathbf{W}$ be a positive definite $n \times n$ matrix. PLS defines κ $\mathbf{W}$-orthogonal *scores* $(\underline{t}_k)_{1 \leq k \leq \kappa}$, linear combinations of the columns of $\mathbf{X}$ and performs a $\mathbf{W}$-weighted least squares regression of $\underline{\mathbf{y}}$ on $(\underline{t}_1, \cdots, \underline{t}_\kappa)$. This yields the decomposition

$$\underline{\mathbf{y}} = q_1 \underline{t}_1 + \cdots + q_\kappa \underline{t}_\kappa + \underline{f}_{\kappa+1} = \mathbf{X} \hat{\underline{\beta}}^{\mathrm{PLS},\kappa} + \underline{f}_{\kappa+1}$$

where the residual term $\underline{f}_{\kappa+1}$ is $\mathbf{W}$-orthogonal to the vectors $(\underline{t}_1, \cdots, \underline{t}_\kappa)$. Note that we do not consider the intercept term in this decomposition since it does not appear in the Cox regression model. Contrary to classical dimension reduction methods (such as Principal Component Regression), the scores depend on the response vector $\underline{\mathbf{y}}$; roughly speaking, given $(\underline{t}_k)_{1 \leq k \leq l}$, $\underline{t}_{l+1}$ is the linear combination of the columns of $\mathbf{X}$, *i.e.* is on the form $\underline{t}_{l+1} = \mathbf{X}\underline{c}$, which is the most informative on the residual response variable $\underline{f}_{l+1}$, when information is defined in terms of the weighted covariance $|\mathrm{Cov}(\sqrt{W}\mathbf{X}c, \sqrt{W}\underline{f}_{l+1})|$ ($\sqrt{\mathbf{W}}$ denotes the square root matrix of $\mathbf{W}$) [9]. While the maximal number of PLS scores $\kappa_{\max}$ can be lower than $\mathrm{rank}(\mathbf{X})$, in practice, it is often equal to $\mathrm{rank}(\mathbf{X})$. Helland [9] shows that the WPLS regression applied

with $\kappa = \kappa_{\max}$ is nothing more than the Weighted Least Squares regression. In the literature, PLS is usually derived with $\mathbf{W} = \mathrm{I}$, the identity matrix; we detail here the algorithm in the weighted case.

1. $\mathbf{E}^0 = \mathbf{X}$; $\underline{f}_0 = \underline{\mathbf{y}}$.
2. For $k = 1, \cdots, \kappa$,

   $\underline{t}_k = \mathbf{E}^{k-1}(\mathbf{E}^{k-1})^T\mathbf{W}\underline{f}_{k-1}$.
   if $k < \kappa$,
   $q_k = \underline{t}_k^T\mathbf{W}\underline{f}_{k-1}/(\underline{t}_k^T\mathbf{W}\underline{t}_k)$,
   $\underline{f}_k = \underline{f}_{k-1} - q_k\underline{t}_k$,
   $\mathbf{E}^k = \mathbf{E}^{k-1} - \underline{t}_k\underline{t}_k^T\mathbf{W}\mathbf{E}^{k-1}/(\underline{t}_k^T\mathbf{W}\underline{t}_k)$.

Hereafter, this procedure is denoted by WPLS $(\underline{\mathbf{y}}, \mathbf{X}, \mathbf{W}, \kappa)$. If $\mathbf{X}$ is full column-rank, this algorithm determines an unique estimate $\hat{\underline{\beta}}^{\mathrm{PLS},\kappa}$ satisfying $\underline{\mathbf{y}} - \underline{f}_{k+1} = \mathbf{X}\hat{\underline{\beta}}^{\mathrm{PLS},\kappa}$; if $\mathbf{X}$ is not full column-rank, the procedure above yields the minimal norm vector among all the vectors verifying $\underline{\mathbf{y}} - \underline{f}_{k+1} = \mathbf{X}\underline{\beta}$.

2.5 Ridge Partial Least Squares for Cox model (RCoxPLS)

A direct application of PLS to Cox regression model seems to be intuitively unappealing because PLS does not handle the censoring aspect of the survival data properly. In order to extend PLS to Cox regression model, we want to replace the censoring data vector of $\underline{\tilde{T}}$ with a pseudo-response variable whose expected value has a linear relationship with the covariates. This extension stands in the same spirit as in Fort and Lambert-Lacroix ([6]). The pseudo-response variable $\underline{z}^\infty$ at convergence of (R)IRLS algorithm verifies this condition and is thus our candidate: it can be written $\underline{z}^\infty = \mathbf{X}\hat{\underline{\beta}}^{\mathrm{R}} + \underline{\varepsilon}$, where, conditionally to $\hat{\underline{\beta}}^{\mathrm{R}}$ being the true value of the parameter, $\underline{\varepsilon}$ is a centered vector of covariance matrix $(\mathbf{W}^\infty)^{-1}$. The main advantage of choosing $\underline{z}^\infty$ instead of, for example, the pseudo-variable at convergence of IRLS - which has the linear structure too- is that this allows the combination of a regularization step and of a dimension reduction step. In addition, this extension is always well-defined: recall indeed that in some cases, the PL maximum estimate does not exist so that the pseudo-variable 'at convergence' of IRLS is of infinite norm.
As a consequence, we propose a new procedure which combines Ridge penalty - the regularization step - when $n > p$ and PLS - the dimension reduction step - and so called Ridge-Cox-PLS (RCoxPLS). Let λ be some positive real constant and κ be some positive integer. RCoxPLS divides in two steps:

1. (a) if $p \leq n$, $(\underline{z}^{\infty}, \mathbf{W}^{\infty}) \longleftarrow \text{IRLS}(\tilde{\underline{T}}, \tilde{\underline{\Delta}}, \tau, \mathbf{X})$;
   (b) if $p > n$, $(\underline{z}^{\infty}, \mathbf{W}^{\infty}) \longleftarrow \text{RIRLS}(\tilde{\underline{T}}, \tilde{\underline{\Delta}}, \tau, \mathbf{X}, \lambda)$;
2. $\hat{\underline{\beta}}^{\text{PLS},\kappa} \longleftarrow \text{WPLS}(\underline{z}^{\infty}, \mathbf{X}, \mathbf{W}^{\infty}, \kappa)$.

The first step builds a continuous response variable $\underline{z}^{\infty}$ for the input of PLS, the "dispersion matrix" of which is $[\mathbf{W}^{\infty}]^{-1}$. This explains the call, in the second step, for a weighted PLS procedure with weight $\mathbf{W}^{\infty}$. RCoxPLS depends on two parameters, λ and κ. They can be selected by cross-validation.

The procedure, presently derived in $\mathbb{R}^p$, can be equivalently derived in $\mathbb{R}^r$ where $r = \text{rank}(\mathbf{X}) \leq n$. To that goal, compute $\mathbf{UDV}^T$ the singular values decomposition (svd) of $\mathbf{X}$ and collect the first r columns of $\mathbf{UD}$ in $\mathbf{X}^{red} = (\mathbf{UD})_{\cdot,1:r}$ so that $\mathbf{X}\underline{\beta} = \mathbf{X}^{red}\underline{\theta}$ for some $\underline{\theta} \in \mathbb{R}^r$. It is readily seen that RCoxPLS (see [8] for the PL with Ridge penalty part and [6] for the PLS part), run by replacing $\mathbf{X}$ by $\mathbf{X}^{red}$ and yields an estimate $\hat{\underline{\theta}}^{\text{PLS},\kappa}$ uniquely related to $\hat{\underline{\beta}}^{\text{PLS},\kappa}$ by $\hat{\underline{\beta}}^{\text{PLS},\kappa} = \mathbf{V}_{\cdot,1:r}\hat{\underline{\theta}}^{\text{PLS},\kappa}$. So when $n << p$, we can replace the huge matrix $\mathbf{X}$ with p columns by the much smaller matrix $\mathbf{X}^{red}$ with n columns, and fit the same model in the smaller space. All aspects of model evaluation, including cross-validation, can be performed in this reduced space. That is of computational importance.

2.6 Comparison with other approaches

2.6.1 Principal components Cox regression

The Principal components Cox regression (PCRCox) uses a principal components analysis to summarize the gene expressions by few linear combinations. These first κ principal components are then included in a multivariate Cox regression model. The parameter κ can be selected by cross-validation. As RCoxPLS, PCRCox is a dimension reduction method but this reduction is achieved without regard to the response variable.

2.6.2 Partial Least Squares approaches

We present two other approaches (see [2] and [12]) to extend PLS to Cox model regression. Both extensions can be regrouped in the following algorithm.

ALGORITHM **1** *1. Put* $\mathbf{E}^1 = \mathbf{X}$ *and compute the regression coefficients* $\hat{\beta}_j^1$ *in the*

Cox regression on $\mathbf{E}^1_{\cdot,j}$ *for each variable* $\mathbf{E}^1_{\cdot,j}$, $j = 1, \ldots, p$. *Compute*

$$\underline{t}_1 = \sum_{j=1}^{p} w_{1,j} \hat{\beta}^1_j \mathbf{E}^1_{\cdot,j},$$

where $w_{1,j}$ *are weights to be precised later.*

2. *For* $k = 2, \cdots, \kappa$, *put* $\mathbf{E}^k = \mathbf{E}^{k-1} - \underline{t}_k \underline{t}_k^T \mathbf{E}^{k-1} / (\underline{t}_k^T \underline{t}_k)$. *For each* $j = 1, \ldots, p$, *compute the regression coefficient* $\hat{\beta}^k_j$ *weighting* $\mathbf{E}^k_{\cdot,j}$ *in the Cox regression on* $\underline{t}_1, \ldots, \underline{t}_{k-1}$, *and* $\mathbf{E}^k_{\cdot,j}$. *Compute*

$$\underline{t}_k = \sum_{j=1}^{p} w_{k,j} \hat{\beta}^k_j \mathbf{E}^k_{\cdot,j},$$

where $w_{k,j}, k = 2, \cdots, \kappa - 1$ *are weights to be precised later.*

Once the κ PLS components are determined, a Cox regression model with covariates $\underline{t}_1, \ldots, \underline{t}_\kappa$ is fitted. Both algorithms differ in the choice of the weights w_{kj}. In [12], algorithm "Partial Cox regression" (PartialCox) consists in applying the Algorithm 1 with

$$w_{kj} = \frac{Var_e(\mathbf{E}^k_{\cdot,j})}{\sum_{m=1}^{p} Var_e(\mathbf{E}^k_{\cdot,m})},$$

where Var_e denotes the empirical variance. Let us note that in [12], the covariables are centered. However, this does not affect the results since the intercept term is not included in the Cox regression model.

In [2] approach, Algorithm 1 is applied with weights defined by

$$w_{kj} = \frac{1}{\left\| \hat{\underline{\beta}}^k \right\|_2}.$$

This procedure is called "PLS-Cox Algorithm" (PLSCox). In [2], another difference with the Algorithm 1 stands in the p Cox regression. At each step, the Cox regressions are computed on $\underline{t}_1, \ldots, \underline{t}_k$, and $\mathbf{X}_{\cdot,j}$ instead of $\underline{t}_1, \ldots, \underline{t}_k$, and $\mathbf{E}^k_{\cdot,j}$. But $\mathbf{E}^k_{\cdot,j}$ is equal to $\mathbf{X}_{\cdot,j}$ minus its orthogonal projection on $\underline{t}_1, \ldots, \underline{t}_k$, so the regression coefficients weighting $\mathbf{X}_{\cdot,j}$ or $\mathbf{E}^k_{\cdot,j}$ are the same.

These both procedures depend on parameter κ that can be selected by cross-validation.

When the number of variables exceeds by far the number of observations, as it is the case with gene expression, these algorithms become computer-intensive and

technical problems may arise. As we have seen, PLS linear regression algorithm is invariant under orthogonal transformation of the $\mathbf{X}$, so PLS based on the $\mathbf{X}^{red}$ (or equivalently on the $\mathbf{X}$ principal components) is equivalent to PLS based on $\mathbf{X}$. That is of computational importance. But this invariance property of PLS linear regression does not hold for these extensions of PLS to Cox model regression. In both papers [2] or [12], the authors propose alternative algorithms by using svd even if these algorithms are not invariant under this transformation. In the next section, we present results of both versions of Algorithm 1, with and without svd. We call svdPartialCox and svdPLSCox these new versions.

Let us note that in the standard case ($p < n$) and in the Gaussian regression, the PLS method leads to the same estimator as the ordinary least squares when the PLS components number is equal to p. We have the analogous result for the Cox regression model when considering RCoxPLS. On the other hand, this property does not stand for the extensions proposed by [2] or [12].

2.6.3 Lasso method

The Cox Lasso method was proposed by [21]. Although the first goal of this method is to select a few covariates among a huge amount, it can also be used as a regularization method. This procedure shrinks the regression coefficients in a similar manner as Ridge regression but uses the absolute values instead of the squared values. The Cox Lasso estimator $\hat{\underline{\beta}}^{Lasso}$ is defined as the maximizer of the penalized likelihood

$$l_n^*(\underline{\beta}) = l_n(\underline{\beta}) - \lambda \sum_{j=1}^{p} |\underline{\beta}_j|,$$

where $\lambda > 0$ is the shrinkage parameter. Penalizing with the absolute values involves that a number of the estimated coefficients will become exactly equal to 0. That means that the Lasso is a variable selection method. The parameter λ can be chosen by minimization of the Bayesian information criterion or by cross-validation. In this study we use the Cox Lasso method and the cross-validation implemented in the R package `penalized`.

Table 1 summarizes the properties of these tested methods.

3 Numerical results

In this section we compare all the previous methods by considering applications to survival prediction from gene expression data. Note that all these methods are not

	RCoxPLS	PartialCox	svdPartialCox	PLSCox
1	1	$p\kappa$	$n\kappa$	$p\kappa$
2	yes	no	no	no
3	yes	no	-	no
	svdPLSCox	RCox	PCRCox	Lasso
1	$n\kappa$	1	1	1
2	no	no	yes	no
3	-	yes	-	no

Table 1: Summary of the properties of the tested methods. 1. How many times does the Cox PL need to be maximized in the procedure ? 2. Does the procedure coincide with the classical Cox estimator in the classical case $n < p$ and when $\kappa = p$? 3. Does the procedure retrieve the same estimator when working on the svd transformation instead of the initial matrix?

invariant by standardization of the design matrix but the impact of the standardization is negligible. In this study we present the results for design matrix without standardization.

3.1 Data

Two real data sets were used to compare the presented procedures. Firstly, the well-known breast cancer data set by Van't Veer et al. [23] was used to compare the methods. Several versions of the data exist, we used those described in [24]. It consists of expressions of 24885 genes of 295 patients. Like in [24], the number of genes has been reduced to 5057 using the Rosetta error model (genes with p-values less than 0.001 in 45 of the 295 samples were removed). The rate of censoring for this data set is 64%.

The second data set is the so-called DLBCL data set presented in [19]. Expressions profile of 7399 genes were collected for each individual of a 240 sample. As in [12], we applied a nearest neighbor technique to estimate the missing values. The missing value is replaced by the average of the 8 nearest neighbors according to the Euclidian distance. The rate of censoring for this data set is 57.5%.

3.2 Assessing prediction methods

There exist many ways to assess the performance of the survival prediction in the Cox model: time dependent ROC curves ([12]), likelihood ratio test statistics and/or

its associated p-value, R^2 criterion and Brier score [24], variance of the martingale residuals [1]. In this paper, we choose to consider

- the R^2 criterion based on deviance $dev = -2(l_n(\hat{\underline{\beta}}) - l_n(\underline{0}))$, and $R^2 = 1 - \exp(dev/n)$
- the variance of the martingales residuals $N_i(\tau) - e^{\underline{X}_i^T \hat{\underline{\beta}}} \hat{\Lambda}(\hat{\underline{\beta}}, \tau)$, where $\hat{\Lambda}(\underline{\beta}, .)$ is the Nelson-Aalen estimator of $\int_0^{\cdot} \lambda$
- the integrated R^2 criterion based on the Brier score defined by

$$\begin{aligned} BS(t) &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{S}(t|\underline{X}_i)^2 I\left(\tilde{T}_i \leq t, \Delta_i = 1\right)}{\hat{G}(\tilde{T}_i)} + \frac{(1 - \hat{S}(t|\underline{X}_i))^2 I\left(\tilde{T}_i > t\right)}{\hat{G}(t)} \right) \\ R^2_{BS}(t) &= 1 - \frac{BS(t)}{BS^0(t)}, \\ iRBS &= \frac{1}{\max \tilde{T}_i} \int_0^{\max \tilde{T}_i} R^2_{BS}(t) dt. \end{aligned}$$

Here, $\hat{G}$ denotes the Kaplan-Meier estimator of the censoring variables U_i and $\hat{S}(t|\underline{X}_i)$ the survival estimator defined by $\hat{S}(t|\underline{X}_i) = \exp(-e^{\underline{X}_i^T \hat{\underline{\beta}}} \Lambda(\hat{\underline{\beta}}, t))$, BS^0 is the Kaplan-Meier estimator based on the $\tilde{T}_i, \Delta_i$ (corresponding to a prediction without covariates). The criterion iRBS has already been used in [3], for instance. Note that $BS(t)$ and $R^2_{BS}(t)$ are piecewise constant functions of the time t. The integral in iRBS is then simply calculated in summing up the product of the values on each interval by the length of each interval.

We perform re-randomization study *i.e.* an out of sample analysis on 100 random subdivisions of the data set into a learning set and a test set. We choose a test set size equal to one third of the data (2:1 scheme of [5]). Each subdivision yields a test set error rate for each predictor; boxplots are used to summarize these error rates over the runs.

3.3 Hyper-parameters choice

The optimal number of PLS or PCR components is selected by choosing the value of κ in the range $\{1, 2, \ldots, 6\}$ by a 5-fold cross validation on each of the 100 training sets. That is, each training set is splited five fold into a test set with size equal to one fifth of the data and a learning set size equal to four fifth of the remaining data. We

retain the value of κ which minimizes the mean of variance of the martingales residuals over these 5-fold cross validation. This is also employed for the regularization parameter λ for RCox for 60 $\log_{10}$-linearly spaced points in the range $[10^{-3}; 10^{4}]$. For RCoxPLS, the κ in $\{1, 2, \ldots, 6\}$ value and λ for 6 $\log_{10}$-linearly spaced points in the range $[10^{-3}; 10^{2}]$ are simultaneously determined by this cross validation method. An alternative would be to choose first the optimal Ridge parameter, and in a second step, the number of components, but this latter procedure appeared to be not very stable. In particular the λ values does not have the same order for RCoxPLS and for RCox. The Lasso parameter is selected by a 5-fold cross validation (by using the `optL1` in the package `penalized`) in the range $[10^{-3}; 10]$ (resp. $[10^{-3}; 70]$) for the first (resp. second) data set.

3.4 Results and discussion

Tables 2 and 3 present means and standard deviations (in brackets) for breast cancer data and for the DLBCL data over the 100 splits of the following quantities. For all methods that use a PLS step and for PCRCox we give the mean number of selected components κ. For Ridge and Lasso type procedures, the mean value of the shrinkage parameters λ is given. The mean CPU time of corresponding to the hyper-parameters research (CPU1) and the one corresponding to the procedure itself (CPU2) are presented. Finally we give the 3 performance indicators for the 8 procedures: RCoxPLS (our), RCox (Ridge Cox without dimension reduction step), PCRCox (Cox fit on PCA components of the design matrix), PartialCox (Li and Gui's approach, with and without svd), PLSCox (Bastien's approach, with and without svd) and Cox Lasso. Since no real effect of the standardization can be observed on the results, we only present here the results without standardization. Figures 1 and 2 give boxplots associated with the 3 performance indicators. Small values of the variance of martingale residuals and high values of R^2 indicators indicate a good performing method.

For the Breast Cancer data set, we can see that around 1 or 2 components are necessary to summarize the information to predict the survival times. Note that RCoxPLS only needs one component on average like PCRCox, whereas the other methods need 2 on average. Note also that RCox usually chooses very high values of λ, whereas RCoxPLS chooses reasonable values. The CPU times show that the hyper-parameters research represents a very large part of time of calculation with respect to the procedures themselves. In particular, this time is quite important for the Ridge parameter (RCox and RCoxPLS). Note also that using the svd improve the time of calculation for PartialCox and PLSCox. In terms of performance, one can see that

Methods	κ	λ	CPU1	CPU2	Residuals	devR2	iRBS
RCoxPLS	1.05	8.55	570.21	14.66	0.62 (0.30)	0.05 (0.08)	0.20 (0.06)
RCox	–	2316.30	639.20	6.23	0.82 (4.65)	0.06 (0.08)	0.21 (0.05)
Lasso	–	6.99	19.66	1.71	0.38 (0.38)	0.06 (0.06)	0.21 (0.05)
PCRCox	1.05	–	94.32	5.60	0.28 (0.06)	0.06 (0.04)	0.20 (0.04)
svdPartialCox	1.99	–	45.67	8.28	0.53 (0.19)	-0.11 (0.14)	0.08 (0.10)
PartialCox	1.92	–	884.68	57.32	0.66 (0.43)	-0.06 (0.15)	0.12 (0.10)
svdPLSCox	1.99	–	46.83	8.13	0.53 (0.19)	-0.11 (0.14)	0.08 (0.10)
PLSCox	1.85	–	960.41	55.58	1.20 (4.71)	-0.13 (0.28)	0.09 (0.13)

Table 2: **Breast Cancer data:** means and standard deviations (in brackets) over the 100 splits of the hyper-parameters, the 3 performance indicators and the CPU times of calculation for 8 procedures: RCoxPLS (our), RCox (Ridge Cox without dimension reduction step), Lasso (Tibshirani's Cox Lasso), PCRCox (Cox fit on PCA components of the design matrix), PartialCox (Li and Gui's approach) with and without svd and PLSCox (Bastien's approach) with and without svd.

Methods	κ	λ	CPU1	CPU2	Residuals	devR2	iRBS
RCoxPLS	1.00	13.75	64.41	7.34	0.96 (0.67)	0.03 (0.08)	0.29 (0.05)
RCox	–	6302.90	399.45	6.09	11.09 (69.07)	0.08 (0.11)	0.33 (0.06)
Lasso	–	26.67	35.05	1.91	0.82 (2.48)	0.04 (0.06)	0.31 (0.05)
PCRCox	1.05	–	111.34	17.66	0.26 (0.02)	-0.01 (0.01)	0.26 (0.05)
svdPartialCox	1.79	–	36.81	7.55	3.71 (3.69)	-0.07 (0.20)	0.25 (0.10)
PartialCox	1.82	–	1344.24	79.02	4.84 (8.40)	-0.06 (0.17)	0.26 (0.10)
svdPLSCox	1.00	–	37.34	6.60	268.81 (541.02)	-0.31 (0.43)	0.25 (0.11)
PLSCox	1.77	–	1365.08	76.75	5.35 (9.58)	-0.06 (0.17)	0.25 (0.10)

Table 3: **DLBCL data:** means and standard deviations (in brackets) over the 100 splits of the hyper-parameters, the 3 performance indicators and the CPU times of calculation for 8 procedures: RCoxPLS (our), RCox (Ridge Cox without dimension reduction step), Lasso (Tibshirani's Cox Lasso), PCRCox (Cox fit on PCA components of the design matrix), PartialCox (Li and Gui's approach) with and without svd and PLSCox (Bastien's approach) with and without svd.

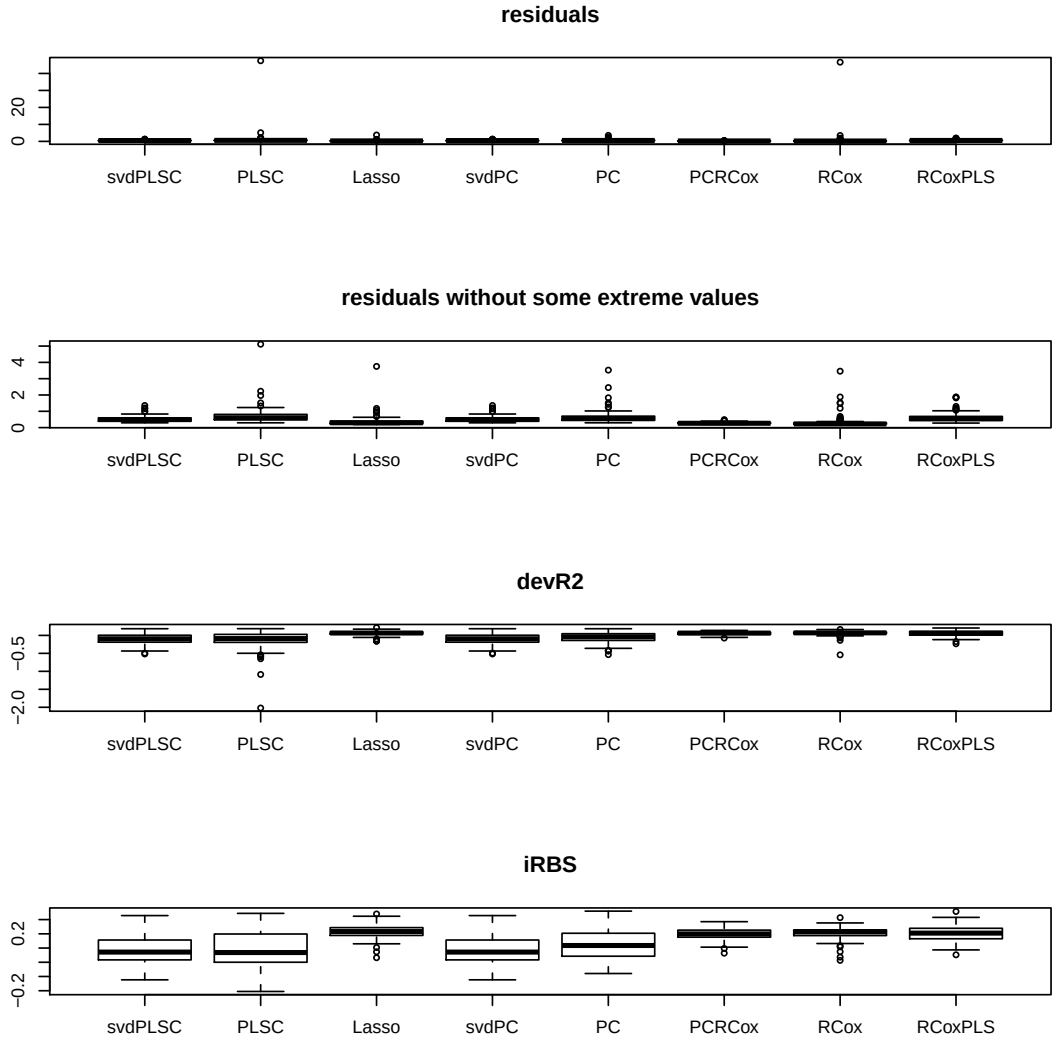

Figure 1: **Breast Cancer data:** box-plots of variance of the martingale residuals, R^2 criterion based on the deviance and integrated R^2 based on the Brier score over 100 cross-validation splits. The second figure is the first one zoomed on, after removing the extreme values of PLSCox and of RCox. (PC for PartialCox and PLSC for PLSCox)

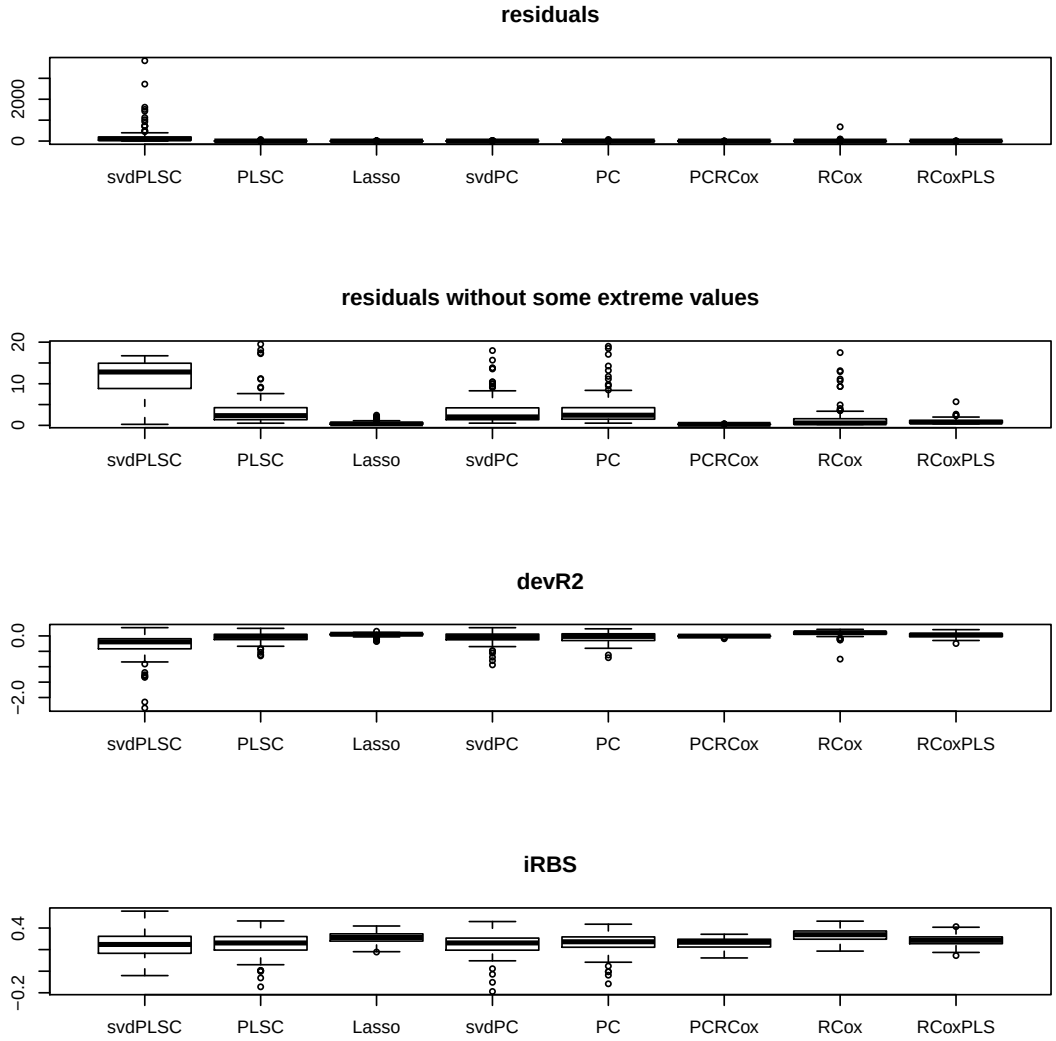

Figure 2: **DLBCL data:** box-plots of variance of the martingale residuals, R^2 criterion based on the deviance and integrated R^2 based on the Brier score over 100 cross-validation splits. The second figure is the first one zoomed on, after removing the extreme values of PLSCox and of Ridge. (PC for PartialCox and PLSC for PLSCox)

PartialCox and PLSCox (with or without svd) produce less stable results, since they produce more outliers and wider boxplots, especially as far as the iRBS indicator is concerned. The RCoxPLS seems to give comparable performance to PCRCox, RCox and Lasso, except as far as the residuals-based indicator is concerned, for which it is slightly worse.

For the DLBCL data set, the results are similar to the first data set in terms of choice of the hyper-parameters, except that PLSCox with svd needs only one component, like RCoxPLS and PCRCox. The CPU time for the research of hyper-parameters in RCoxPLS is now better than both RCox and PCRCox. One can see that PartialCox and PLSCox (with or without svd) perform poorly with respect to the other methods, especially as far as the variance of the martingale residuals is concerned. RCox seems again to produce a lot of outliers for this indicator, which calls for a dimension reduction step. For this data set also, RCoxPLS performs roughly as well as PCRCox and Lasso.

4 Conclusion

In this work, we compared 8 dimension reduction and/or regularization methods for survival data in the presence of a great amount of regressors with respect to the number of individuals, a case which typically arises in gene expression data. In particular, we present a combination of a Ridge Cox regression and a PLS procedure that appears to compete quite well with the previous PLS methods. More precisely, the latter method seems better in performance than PartialCox and PLSCox, two methods adapting the PLS algorithm for the Cox model. RCoxPLS performs comparably to PCRCox and RCox, but it is faster to compute in some cases (in the DLBCL case, for instance). Note furthermore, that RCox produces a lot of extreme values, probably due to a too high value of its hyper-parameter, choosen by cross-validation.

As mentioned by a referee, an interesting problem linked to this one would be to consider the added value of gene expression data with respect to clinical ones. Nevertheless, this would require to consider an iterative procedure like suggested by [11] in the Gaussian case, and it is not clear whether it will be possible for the PartialCox and PLSCox procedures. Therefore, a great work would be needed, that is out of the scope of this paper and that we left for future research.

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24.

References

[1] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer-Verlag, New York, 1993.

[2] P. Bastien. PLS-Cox model: application to gene expression. In *COMPSTAT 2004—Proceedings in Computational Statistics*, pages 655–662. Physica, Heidelberg, 2004.

[3] H. M. Bovelstad and Ø. Borgan. Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal*, 53:202–216, 2011.

[4] D. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B*, 74:187–220, 1972.

[5] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, 97:77–87, 2002.

[6] G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104–1111, 2005.

[7] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools,with discussion. *Technometrics*, 35(2):109–148, 1993.

[8] T. Hastie and R. Tibshirani. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5(3):329–340, 2004.

[9] I. Helland. On the structure of Partial Least Squares Regression. *Commun. Stat., Simulation Comput.*, 17(2):581–607, 1988.

[10] P. J. Green Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of the Royal Statistical Society, B*, 46(2):149–192, 1984.

[11] K. Jorgensen, V. H. Segtnan, K. Thyholt, and T. Naes. A Comparison of Methods for Analysing Regression Models with Both Spectral and Designed Variables. *Journal of Chemometrics*, 18(10):451-464, 2004.

[12] H. Li and J. Gui. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(1):i208–i215, 2004.

[13] B. D. Marx. Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression. *Technometrics*, 38(4):374–381, 1996.

[14] W. F. Massy. Principal components regression in exploratory statistical research. *J. Amer. Stat. Assoc.*, 60:234–246, 1965.

[15] T. Naes and H. Martens. Comparison of prediction methods for multicollinear data. *Commun. Stat., Simulation Comput.*, 14:545–576, 1985.

[16] D. Nguyen and D. Rocke. Tumor classification by Partial Least Squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[17] S. Nygard, Ø. Borgan, O. C. Lingjaerde and H. L. Storvold Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, 14:179–195, 2008.

[18] P. J. Park, L. Tian, and I. S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18:S120–S127, 2002.

[19] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, L. Yang, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. LÛpez-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, , and L. M. Staudt. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New English Journal of Medicine*, 346:1937–1947, 2002.

[20] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. -L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA.*, 100(14):8418–8423, 2003.

[21] R. Tibshirani. The Lasso method for variable selection in the Cox model. *Stat. Med.*, 16:385–395, 1997.

[22] H. van Houwelingen, T. Bruinsma, H. A.A.M, van't Veer L.J., and L. Wessels. Cross-validated Cox regression on microarray gene expression data. *Stat. Med.*, 25:3201–3216, 2006.

[23] L. J. van Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[24] W. N. van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix. Survival prediction using gene expression data: A review and comparison. *Computational Statistics & Data Analysis*, 53(5):1590–1603, 2009.

[25] H. Wold. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In *Perspect. Probab. Stat., Pap. Honour M. S. Bartlett Occas. 65th Birthday*, pages 117–142, 1975.