BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 3 2015, pages 397–404
doi:10.1093/bioinformatics/btu660

Data and text mining

Advance Access publication October 6, 2014

Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data

Philippe Bastien[1,*,†], Frédéric Bertrand[2,*,†], Nicolas Meyer[3] and Myriam Maumy-Bertrand[2]

[1]L'Oréal Recherche & Innovation, 93601 Aulnay-sous-Bois, [2]IRMA, CNRS UMR 7501, Labex IRMIA, Université de Strasbourg, 67084 Strasbourg Cedex, [3]INSERM EA3430, Laboratoire de Biostatistique, Faculté de Médecine de Strasbourg, Labex IRMIA, Université de Strasbourg, 67085 Strasbourg Cedex, France

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: A vast literature from the past decade is devoted to relating gene profiles and subject survival or time to cancer recurrence. Biomarker discovery from high-dimensional data, such as transcriptomic or single nucleotide polymorphism profiles, is a major challenge in the search for more precise diagnoses. The proportional hazard regression model suggested by Cox (1972), to study the relationship between the time to event and a set of covariates in the presence of censoring is the most commonly used model for the analysis of survival data. However, like multivariate regression, it supposes that more observations than variables, complete data, and not strongly correlated variables are available. In practice, when dealing with high-dimensional data, these constraints are crippling. Collinearity gives rise to issues of over-fitting and model misidentification. Variable selection can improve the estimation accuracy by effectively identifying the subset of relevant predictors and enhance the model interpretability with parsimonious representation. To deal with both collinearity and variable selection issues, many methods based on least absolute shrinkage and selection operator penalized Cox proportional hazards have been proposed since the reference paper of Tibshirani. Regularization could also be performed using dimension reduction as is the case with partial least squares (PLS) regression. We propose two original algorithms named sPLSDR and its non-linear kernel counterpart DKsPLSDR, by using sparse PLS regression (sPLS) based on deviance residuals. We compared their predicting performance with state-of-the-art algorithms on both simulated and real reference benchmark datasets.

Results: sPLSDR and DKsPLSDR compare favorably with other methods in their computational time, prediction and selectivity, as indicated by results based on benchmark datasets. Moreover, in the framework of PLS regression, they feature other useful tools, including biplots representation, or the ability to deal with missing data. Therefore, we view them as a useful addition to the toolbox of estimation and prediction methods for the widely used Cox's model in the high-dimensional and low-sample size settings.

Availability and implementation: The R-package plsRcox is available on the CRAN and is maintained by Frédéric Bertrand. http://cran.r-project.org/web/packages/plsRcox/index.html.

Contact: pbastien@rd.loreal.com or fbertran@math.unistra.fr.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 29, 2013; revised on September 13, 2014; accepted on October 1, 2014

1 INTRODUCTION

1.1 Lasso penalized Cox proportional hazards regression

L_1 penalized Cox regression was first suggested by Tibshirani, 1997, as an extension of its variable selection procedure called least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996). Since then many developments in the framework of generalized Lasso have appeared. Efron *et al.*, 2004, described a highly efficient procedure called LARS for Least Angle Regression for variable Selection, which can be slightly modified to provide solution for the Lasso. Gui and Li, 2005, taking advantage of the close connection between LARS and Lasso proposed a solution to the Cox–Lasso procedure in the setting of high-dimensional data, combining LARS path inside each Newton approximation. However, the iterative reweighted least squares strategy used by the LARS–Cox procedure for handling survival endpoints undo much of the computational efficiency of the LARS–Lasso procedure. Segal, 2006, described an estimation strategy that restores the computational efficiency of the LARS–Lasso procedure. He showed that the expression to be minimized in the Cox–Lasso procedure of Tibshirani could be approached by the deviance residuals sum of squares. He then performed the original LARS-Lasso algorithm on those specific residuals. Park and Hastie, 2007, using results from the LARS algorithm, proposed a L_1-regularization path algorithm for generalized linear models based on the predictor-corrector method of convex optimization that they applied to the Cox proportional hazards model. Their algorithm exploits the near piece-wise linearity of the coefficients to approximate the solution at different constraints, then numerically maximizes the likelihood for each constraint via a Newton iteration initialized at the approximation. Sohn *et al.*, 2009, extended the Kim *et al.*, 2008, coordinate-wise gradient Lasso algorithm to the Cox proportional hazard model. Following quite the same ideas, Goeman, 2010, described a full gradient Lasso algorithm. Unlike the algorithm of Sohn *et al.*, it does not update a single coordinate at a time, but uses the full gradient at each step and can switch to a Newton–Raphson algorithm to speedup its convergence when it gets close to the optimum. Both methods are gradient based and therefore scalable to high-dimensional data because they do not require matrix inversion. Tibshirani, 2009, proposed *uniCox*,

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

a Cox univariate shrinkage method, assuming that the features are independent in each risk set, in the spirit of the univariate soft thresholding solution for the Lasso in linear regression when the features are independent. More recently, Simon *et al.*, 2013, proposed *coxnet*, a fast path-wise algorithm for the Cox's model regularized by convex combination of L_1 and L_2 penalties, named *Elastic Net* by Zou and Hastie, 2005. Their algorithm uses cyclical coordinate descent extending the work of Friedman *et al.*, 2010, with early reference from the shooting algorithm of Fu, 1998. Other, though non-exhaustive, regularized Cox regression procedures include the smoothly clipped absolute deviation of Fan and Li, 2002, the adaptive Lasso of Zhang and Lu, 2007, the Dantzig selector of Antoniadis *et al.*, 2010, the Sure and Iterative Sure Independence Screening of Fan *et al.* (2010) or the hierarchically penalized Cox regression of Wang *et al.*, 2009, which yields sparsity at both group and individual feature levels to select groups and predictors within a group, for example, genes within a pathway.

1.2 PLS regression

Prediction in high-dimensional and low-sample size settings already arose in chemistry in the eighties. Partial least squares (PLS) regression, that can be viewed as a regularization method based on dimension reduction, was developed as a chemometric tool in an attempt to find reliable predictive models with spectral data (Tenenhaus, 1998; Wold *et al.*, 1983). Nowadays, the difficulty encountered with the use of genomic or proteomic data for classification or prediction, using large matrices, is of comparable nature. It was thus natural to use PLS regression principles in this new context. The method starts by constructing latent components, using linear combinations of the original variables, which are then used as new descriptors in standard regression analysis. Different from the principal components analysis (PCA), this method makes use of the response variable in constructing the latent components. The PLS regression can be viewed as a regularized approach searching the solution in a subspace named Krylov space giving biased regression coefficients but with lower variance. In the framework of censored genomic data, the PLS regression operates a reduction of the dimensionality of the gene's space oriented toward the explanation of the hazard function. It allows transcriptomic signatures correlated to survival to be determined.

1.3 PLS–Cox regression

Garthwaite, 1994, showed that PLS regression could be obtained as a succession of simple and multiple linear regressions. Tenenhaus, 1999, proposed a fairly similar approach but one that could cope with missing data by using the principles of the Nipals algorithm (Wold, 1966). As a result, Tenenhaus suggested that PLS regression be extended to logistic regression (PLS–LR) by replacing the succession of simple and multiple regressions by a succession of simple and multiple logistic regressions in an approach much simpler than that developed by Marx, 1996. By using this alternative formulation of the PLS regression, Bastien and Tenenhaus, 2001, extended the PLS regression to any generalized linear regression model (PLS–GLR) and to the Cox model (PLS–Cox) as a special case. Further improvements have then been described (Bastien *et al.*, 2005) in the case of categorical descriptors with model validation by bootstrap resampling and variable selection using hard thresholding. Since then many developments in the framework of PLS and Cox regressions have appeared in the literature. Nguyen and Rocke, 2002, directly applied PLS regression to survival data and used the resulting PLS components in the Cox model for predicting survival time. However, such a direct application did not really generalize PLS regression to censored survival data, as it did not take into account the failure time in the dimension reduction step. Based on a straightforward generalization of Garthwaite, 1994, presented a solution, partial Cox regression, quite similar to the one proposed by Bastien and Tenenhaus, using different weights to derive the PLS components but not coping with missing data.

1.4 PLSDR

Following Segal, 2006, who suggested initially computing the null deviance residuals and then using these as outcomes for the LARS-Lasso algorithm, Bastien, 2008, proposed PLSDR, an alternative in high-dimensional settings using deviance residuals-based PLS regression. This approach is advantageous by both its simplicity and its efficiency because it only needs to carry out null deviance residuals using a simple Cox model without covariates and use these as outcome in a standard PLS regression. The final Cox model is then carried out on the *m*-retained PLSDR components.

Moreover, following the principles of the Nipals algorithm, weights, loadings and PLS components are computed as regression slopes. These slopes may be computed even when there are missing data: let $t_{hi} = x_{h_1,i} w_h / w'_h w_h$ the value of the PLS component for individual i, with descriptors matrix X and covariance weights matrix W, t_{hi} represents the slope of the OLS line without constant term related to the cloud of points $(w_h, x_{h_1,i})$. In such case, in computing the h^{th} PLS component, the denominator is computed only on the data available also for the denominator.

1.5 sPLS

Recently, Chun and Keles, 2010, provided both empirical and theoretical results showing that the performance of PLS regression was ultimately affected by the large number of predictors. In particular, a higher number of irrelevant variables leads to inconsistent coefficient estimates in the linear regression setting. There is a need to filter the descriptors as a preprocessing step before PLS fit. However, commonly used variables filtering approaches are all univariate and ignore correlation between variables. To solve these issues, Chun and Keles proposed 'sparse PLS regression', which promotes variables selection within the course of PLS dimension reduction. sPLS has the ability to include variables that variable filtering would select in the construction of the first direction vector. Moreover, it can select additional variables, i.e. variables that become significant once the response is adjusted for other variables in the construction of the subsequent direction vectors. This is the case of 'proxy genes' acting as suppressor variables that do not predict the outcome variable directly but improve the overall prediction by enhancing the effects of prime genes despite having no direct predictive power, Magidson and Wassmann, 2010.

A direct extension of PLS regression to sPLS regression could be provided by imposing L_1 constraint on PLS direction vector w:

$$\max_{w} w'Mw \text{ subject to } w'w=||w||_2=1,\ ||w||_1 \leq \lambda,$$

$$\text{where } M=X'YY'X.$$

When $Y = X$, the objective function coincides with that of sPCA (Jolliffe et al., 2003). However, in that case, Jolliffe *et al.* pointed out that the solution tends not to be sparse enough and the problem is not convex. To solve these issues, Chun and Keles provided an efficient implementation of sPLS based on the LARS algorithm by generalizing the regression formulation of sPCA of Zou et al., 2006:

$$\min_{w,c} \ -\kappa w'Mw+(1-\kappa)(c-w)'M(c-w)+\lambda_1||c||_1+\lambda_2||c||_2$$

$$\text{subject to } w'w=1, \text{ where } M=X'YY'X.$$

This formulation promotes exact zero property by imposing L_1 penalty onto a surrogate of the direction vector c instead of the original direction w while keeping w and c close to each other. The L_2 penalty takes care of the potential singularity of M. Moreover, they demonstrated that for univariate PLS, y regressed on X, the first direction vector of the sparse PLS algorithm was obtained by soft thresholding of the original PLS direction vector:

$$\left(|Z|-\frac{\lambda}{2}\right)+\text{sign}(Z), \text{ where } Z=X'y/||X'y||_2.$$

To inherit the property of the Krylov subsequences, which is known to be crucial for the convergence of the algorithm (Krämer, 2007), the thresholding phase is followed by a PLS regression on the previously selected variables. The algorithm is then iterated with y replaced by $y-X\hat{\beta}$, the residuals of the PLS regression based on the variables selected from the previous steps. The sPLS algorithm leads therefore to sparse solutions by keeping the Krylov subsequence structure of the direction vectors in a restricted X space, which is composed of the selected variables. The thresholding parameter λ and the number of hidden components are tuned by cross-validation.

sPLS has connections to other variable selection algorithms including the elastic net method (Zou and Hastie, 2005) and the threshold gradient method (Friedman and Popescu, 2004). The elastic net algorithm deals with the collinearity issue in variable selection problem by incorporating the ridge regression method into the LARS algorithm. In a way, sPLS handles the same issue by fusing the PLS technique into the LARS algorithm. sPLS can also be related to the threshold gradient method in that both algorithms use only thresholded gradient and not the Hessian. However, sPLS achieves fast convergence by using conjugate gradient. Hence, LARS and sPLS algorithms use the same criterion to select active variables in the univariate case. However, the sPLS algorithm differs from LARS in that sPLS selects more than one variable at a time and uses the conjugate gradient method to compute coefficients at each step. The computational cost for computing coefficients at each step of the sPLS algorithm is less than or equal to the computational cost of computing step size in LARS, as conjugate gradient methods avoid matrix inversion.

1.6 sPLSDR and DKsPLSDR

In this article, we propose two new algorithms, named sPLSDR and DKsPLSDR, by using sPLS or its non-linear kernel counterpart DKsPLS instead of PLS in the PLSDR algorithm. We show them as efficient sparse regularized alternatives based on dimension reduction.

2 METHODS

2.1 The Cox proportional hazards model

The model assumes the following hazard function for the occurrence of an event at time t in the presence of censoring:

$$\lambda(t)=\lambda_0(t)\exp(\beta'X),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, β the vector of the regression coefficients and $X=(X_1,\ldots,X_p)$ a n by p matrix of features with $x_i=(x_{i1},\ldots,x_{ip})$ the covariate vector for the i^{th} individual. The event could be death or cancer relapse. Based on the available data, the Cox's partial likelihood can be written as follows:

$$PL(\beta)=\prod_{k\in D}\frac{\exp(\beta'x_k)}{\sum_{j\in R_k}\exp(\beta'x_j)},$$

where D is the set of indices of the events and R_k denotes the set of indices of the individuals at risk at time t_k.

The goal is to find the coefficients $\hat{\beta}$, which maximize the log partial likelihood function

$$l(\beta)=\log PL(\beta).$$

The vector $\hat{\beta}$ is the solution of the equation:

$$u(\beta)=\frac{\partial l}{\partial\beta}=0$$

with $u(\beta)$ the vector of efficient scores.

However, there is no explicit solution and the minimization is generally accomplished using the Newton–Raphson procedure. An estimate of the vector of β parameters at the $(k+1)^{\text{th}}$ cycle of the iterative procedure is as follows:

$$\hat{\beta}_{k+1}=\hat{\beta}_k+I^{-1}(\hat{\beta}_k)u(\hat{\beta}_k)$$

where $I(\beta)=-\frac{\partial^2 l}{\partial\beta\partial\beta'}$ is the observed information matrix. The process can be started by taking $\hat{\beta}_0=0$ and iterated up to convergence, i.e. when the change in the log likelihood function is small enough. When the iterative procedure has converged, the variance-covariance matrix of the parameter estimates can be approximated by the inverse of the observed information matrix $I^{-1}(\hat{\beta})$.

When $p>n$, there is no unique $\hat{\beta}$ to maximize this log partial likelihood function. Even when $p\leq n$, covariates could be highly correlated and regularization may still be required to reduce the variances of the estimates and to improve the predictive performance.

2.2 Deviance residuals

For the Cox model with no time-dependent explanatory variables and at most one event per patient, the martingale residuals for the i^{th} subject with observation time t_i and event status δ_i, where $\delta_i=0$ if t_i is a censored time, and $\delta_i = 1$ otherwise is as follows:

$$\hat{M}_i = \delta_i - \hat{E}_i = \delta_i - \hat{\Delta}_0(t_i) \exp(\hat{\beta}' x_i)$$

with $\hat{\Delta}_0(t_i)$ the estimated cumulative hazard function at time t_i.

Martingale residuals are highly skewed. The deviance residuals d_i are a normalized transform of the martingale residuals. For the Cox model, the deviance residuals (Collett, 1994) amount to the form:

$$d_i = \text{sign}(\hat{M}_i) \cdot \left[2\left\{-\hat{M}_i - \delta_i \log\left(\delta_i - \hat{M}_i\right)\right\}\right]^{1/2}.$$

The sign function is to ensure that the deviance residuals have the same sign as the martingale residuals. Martingale residuals take values between $-\infty$ and 1. The square root shrinks large negative martingale residuals, while the logarithmic transformation expands toward $+\infty$ martingale residuals that are close to 1. As such, the deviance residuals are more symmetrically distributed around zero than the martingale residuals. The deviance residual is a measure of excess of death and can therefore be interpreted as a measure of hazard. Moreover, Segal showed that the expression to be minimized in step 3 of the Cox–Lasso procedure of Tibshirani can be approximated, in a first order Taylor-series approximation sense, by the deviance residual sum of squares: $(z - X\beta)' A(z - X\beta) \approx \text{RSS}(\hat{D})$ with $\eta = \beta' X$, $\mu = \frac{\partial l}{\partial \eta}$, $A = -\frac{\partial^2 l}{\partial \eta \eta'}$, and $z = \eta + A^{-} \mu$:

2.3 The sPLSDR algorithm

The sPLSDR algorithm involves the following steps:

1. Cox model without covariates to derive the null deviance residuals d.
2. Computation of the sPLS components by using the sPLS regression with the null deviance residuals as outcome.
   a. Set $\hat{\beta}^{PLS} = 0,\ = \{\}$, $k = 1$, $y_1 = d$.
   b. While $(k \leq K)$.
      (1) $w = (|z| - \lambda/2)_+ \text{sign}(z)$ where $z = X' y_1 / ||X' y_1||_2$.
      (2) Update Ω as $\{i : \hat{w}_i \neq 0\} \cup \{i : \hat{\beta}_i^{PLS} \neq 0\}$
      (3) Fit PLS with X by using the k number of latent components.
      (4) Update $\hat{\beta}^{PLS}$ by using the new PLS estimates of the direction vectors and update y_1 and k through $y_1 \leftarrow y_1 - X\hat{\beta}^{PLS}$ and $k \leftarrow k + 1$.
3. Cox model on the m-retained sPLSDR components.

2.4 The DKsPLSDR algorithm

In the case of very many descriptors, PLS regression being invariant by orthogonal transformation (De Jong and ter Braak, 1994), an even faster procedure could be derived by replacing the X matrix by the matrix of principal components Z $(XX' = ZZ')$. This could be viewed as the simple form of linear kernel PLS regression algorithms, which have been proposed in the nineties (Lindgren *et al.*, 1993; Rännar *et al.*, 1994) to solve computational problems posed by large matrices in chemometrics. The objective of these methods was to obtain PLS components by working on a condensed matrix of a considerably smaller size than the original one. Moreover, in addition to dramatically reducing the size of the problem, non-linear pattern in the data could also be analyzed using non-linear kernel.

Rosipal and Trejo, 2001, proposed a non-linear extension of PLS regression using kernels. Assuming a non-linear transformation of the input variables $\{x_i\}_{i=1}^n$ into a feature space F, i.e. a mapping $\Phi : x_i \in \mathbb{R}^N \mapsto (x_i) \in F$, their goal was to construct a linear PLS regression model in F. They derived an algorithm named KPLS for Kernel PLS by performing the PLS regression on $\Phi(X)$. It amounts to replacing, in the expression of PLS components, the product XX' by $\Phi(X)\Phi(X)'$ using the so-called kernel trick, which allows the computation of dot products in high-dimensional feature spaces using simple functions defined on pairs of input patterns: $\Phi(x_i)\Phi(x_j)' = K(x_i, x_j)$. This avoids having to explicitly calculate the coordinates in the feature space, which could be difficult for a highly dimensional feature space. By using the kernel functions corresponding to the canonical dot product in the feature space, non-linear optimization can be avoided and simple linear algebra can be used.

Bennett and Embrecht, 2003, proposed to perform PLS regression directly on the kernel matrix K instead of $\Phi(X)$. DKPLS corresponds to a low rank approximation of the kernel matrix. Moreover, Tenenhaus *et al.* (2007) demonstrated that, for one-dimensional output response, PLS of $\Phi(X)$ (KPLS) is equivalent to PLS on $K^{1/2}$ (DKPLS).

Using previous works, it becomes straightforward to derive a non-linear Kernel sPLSDR algorithm by replacing in the sPLSDR algorithm the X matrix by a kernel matrix K. The main kernel functions are the linear kernel $(K(u, v) = <u, v>)$ and the Gaussian kernel $(K(u, v) = \exp(-||u - v||_2^2 / 2\sigma^2))$.

However, non-linear kernel (sparse) PLS regression loses the explanation with the original descriptors unlike linear kernel PLS regression, which could limit the interpretation of the results.

The DKsPLSDR algorithm involves the following steps:

1. Computation of the kernel matrix.
2. Cox model without covariates to derive the null deviance residuals.
3. Computation of the PLS components by using the DKsPLS algorithm with the null deviance residuals as outcome.
4. Cox model on the m-retained DKsPLSDR components.

3 BENCHMARKING

3.1 Implementation

We benchmarked the new sPLSDR and DKsPLSDR algorithms against the following existing ones: coxpath (Park and Hastie, 2007), coxnet (Simon *et al.*, 2011), PLS-Cox (Bastien and Tenenhaus, 2001), autoPLS-Cox (PLS-Cox with a hard-thresholding approach and automatic selection of the maximal number of components, Bastien *et al.*, 2005), LARS-LassoDR (Segal, 2006), Cox-PLS (Nguyen and Rocke, 2002), PLSDR (Bastien, 2008), DKPLSDR (Bastien, 2008), uniCox (Tibshirani, 2009) and glcoxph (Sohn *et al.*, 2009).

More insights on the implementation of these algorithms are given in the Supplementary Information. We made several wrappers for the Cox-PLS, LARS-LassoDR, PLSDR, sPLSDR, DKPLSDR and DKsPLSDR and had to implement in the R language PLS-Cox, and hence autoPLS-Cox, which is PLS-Cox with hard thresholding of the non-significant explanatory variables at a given α level.

3.2 Prediction evaluation

We propose to use the time-dependent receiver-operator characteristics (*ROC*) curve for censored data (Heagerty and Zheng, 2005) to assess how well the model predicts survival. Let sensitivity and specificity be the following:

$$\text{sensitivity}\,(c, t/X\beta) = \mathbb{P}(X\beta > c/\delta(t) = 1)$$

$$\text{specificity}\,(c, t/X\beta) = \mathbb{P}(X\beta \leq c/\delta(t) = 0)$$

Table 1. Datasets' structure

Nbr.	Dataset	Number of rows	Number of columns	Pct. uncensored
1	Alizadeh	40	4026	45.0
2	Beer	86	7129	72.1
3	Bhattacharjee	125	3171	42.4
4	Garber	22	3171	36.4
5	Metzeler	242	44 754	38.0
6	Romain	117	39	77.8
7	Rosenwald	240	7399	42.5
8	Wang	286	22 283	62.6

with $X\beta$ a predictor score function and $\delta(t)$ the event indicator at time t. Sensitivity measures the expected fraction of subjects with a marker greater than c among the subpopulation of individuals who die at time t, whereas specificity measures the fraction of subjects with a marker less than or equal to c among those who survive beyond time t.

Using the true- and false-positive rate functions, $\mathrm{TP_t}(c) = \text{sensitivity}(c, t)$ and $\mathrm{FP_t}(c) = 1 - \text{specificity}(c, t)$ allows the ROC curve to be written as follows:

$$\mathrm{ROC_t}(p) = \mathrm{TP_t}((\mathrm{FP_t})^{-1}(p)),$$

$$\text{with } (\mathrm{FP_t})^{-1}(p) = \inf_c \{c : \mathrm{FP_t}(c) \leq p\}.$$

The area under the ROC curves (AUC), which measures the probability that a marker value for a randomly selected case exceeds the marker value for a randomly selected control, is particularly useful for comparing the discriminatory capacity of different potential biomarkers. A larger AUC at time t based on the risk score function $X\beta$ indicates better predictability of time to event at time t as measured by sensitivity and specificity at time t.

A typical complexity with survival data is that observations may be censored. Two ROC curve estimators are proposed that can accommodate censored data (Heagerty *et al.*, 2000). A simple estimator is based on using the Kaplan–Meier (KM) estimator for each possible subset $X\beta > c$. However, this estimator does not guarantee the necessary condition that sensitivity and specificity are monotone in X. An alternative estimator that does guarantee monotonicity is based on a nearest-neighbor estimator (NNE) for the bivariate distribution function of (X, T), where T represents survival time (Akritas, 1994). Moreover, it is a semiparametric efficient estimator and the censoring process is allowed to depend on the diagnostic marker X, whereas the Kaplan–Meier estimator assumes that the censoring process does not depend on X (Akritas, 1994). We provide benchmark results for both the KM and NNE estimators.

3.3 Datasets

3.3.1 Simulated datasets We performed a simulation study (summed up in Supplementary Table S1) to evaluate the methods by simulating 100 datasets with exponential survival distribution and 40% censored rate (100 observations × 1000 genes) according to three different simulation types [cluster by Bair *et al.* (2006), factorial by Kaiser and Dickman (1962) and Fan *et al.* (2002) or eigengene by Langfelder *et al.* (2013)], using either no link or a linear one between the response and the predictors.

We divided each of these 600 datasets into a learning set, of 7/10 (70) of the observations, used for estimation, and a test set, of 3/10 (30) of the observations, used for evaluation or testing of the prediction capability of the estimated model. This choice was made to stay between the 2:1 scheme of Bøvelstad *et al.* (2007); Lambert-Lacroix and Letué (2011); van Wieringen *et al.* (2009); and the 9:1 scheme of Li (2006). The division between learning and test sets was balanced using the caret package according to both the response value and censor rate.

3.3.2 Real datasets Eight datasets were used to benchmark the models: Alizadeh *et al.*, 2000; Beer *et al.*, 2002; Bhattacharjee *et al.*, 2001; Garber *et al.*, 2001; Metzeler *et al.*, 2008; Romain *et al.*, 2010; Rosenwald *et al.*, 2002 and Wang *et al.*, 2005. The variability of their numbers of subjects and variables, see Table 1, makes their use as benchmarks more insightful. The allelotyping set of Romain *et al.*, 2010, is a new benchmark set, whereas the other seven are commonly used ones.

4 RESULTS

4.1 Simulated datasets

Cross-validation techniques [either CV partial likelihood criterion (CVLL) or van Houwelingen CV partial likelihood criterion (vHCVLL)] were recommended for several of our benchmark methods by their authors. We followed these recommendations. For the Cox-PLS, PLSDR, sPLSDR, DKPLSDR and DKsPLSDR, we used the iAUCSurvROC criterion, for the LARS-LassoDR the vHCVLL and for PLS-Cox and autoPLS-Cox the iAUCSH. Simulation results confirmed these as relevant cross-validation criteria (Supplementary Figs S3 and S4).

To assess the goodness of fit and prediction accuracy of all the methods, we selected four indices of various kind: the coefficient (R^2XO) proposed by Xu and O'Quigley (1999) (a R^2-like measure and a likelihood-based approach), the *CGH* by Gonen and Heller (2008) (an improved version of the *C* index Harrell *et al.,* 1996), the *iAUCsurvROC* (a ROC-based approach used by Li, 2006) and the integrated weighted Schmid Score (iSSw, Schmid *et al.*, 2011, an integrated robust prediction error and a distance-based approach). The simulations lead to some advantage for the models featuring components (Fig. 1 and Supplementary Figs S6–S8).

4.2 Real datasets

For every dataset and every method, we selected the parameters by applying the same cross-validation techniques as in Section 4.1. The test and train groups were commonly used ones whenever available or else randomly chosen ones. All the scripts used to perform these analyses will be provided as demos in an updated version of the plsRcox package. The datasets will be available as direct downloads on a dedicated academic Web site.

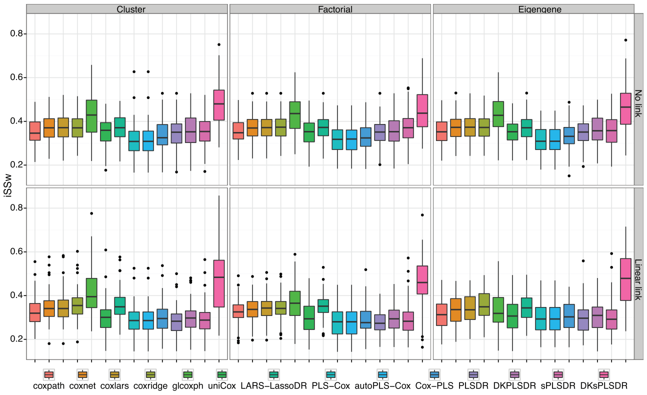

Fig. 1. Methods performance evaluation according to the integrated weighted Schmid Score, a distance-based integrated robust prediction error

We compared the survival prediction accuracy, using the *iAUCsurvROC* criterion, of all the methods over two periods starting from 0 and with two different (near or far) end times. These were chosen separately for every dataset as the first quartile (Q_1) and the third quartile (Q_3) of the empirical distribution of the survival time. This is tantamount to finding models for making accurate survival prognostics either at an earlier or at a later time.

The results are displayed on graphical outputs to point out which is the best method for predicting survival. For instance, the graphical output for comparing the methods' accuracy for predicting the survival at an early time ($[0, Q_1]$ interval) for the Dataset 8 of Wang *et al.*, 2005 is shown on Figure 2. As for survival prediction at a later time ($[0, Q_3]$ interval), results for Dataset 5 of Metzeler *et al.*, 2008 are displayed on Figure 3. All of them show better predictive performance for both sPLSDR and DKsPLSDR algorithms. All graphical outputs are given in Supplementary information (Supplementary Figs S11–S26.

We summed up all the results (Supplementary Tables S3–S6 in Tables 2 and 3 where the methods are ranked by decreasing iAUC. We used **bold** fonts for the new sPLSDR and DKsPLSDR algorithms and *italic* for the first R implementation of existing techniques (Cox-PLS, PLS-Cox, autoPLS-Cox, PLSDR, DKPLSDR) by Bastien and Tenenhaus (2001) and Bastien (2008). These results show steady better predictive performance for the sPLSDR and DKsPLSDR algorithms as well as for the uniCox one. Yet, uniCox failed on the Dataset 2 (Beer *et al.*, 2002), for both $[0, Q_1]$ and $[0, Q_3]$ intervals, even after several tries with random training sets.

Variable selection results were analyzed for the Romain *et al.*, 2010 dataset (Supplementary Table S11, and Supplementary Figs S26 and S27). In our view, variable selection is a topic in itself and especially its robustness with respect to several criteria including censorship or missingness patterns as well as noise or resampling. We will focus on these issues in a subsequent article.

5 DISCUSSION

Overall, DKsPLSDR and, even more, sPLSDR compare favorably with the benchmark methods on both simulated and real datasets.

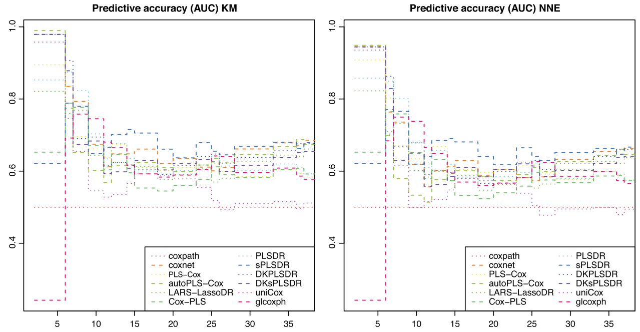

Fig. 2. Dataset 8: Survival prediction over a short period ($[0, Q_1]$)

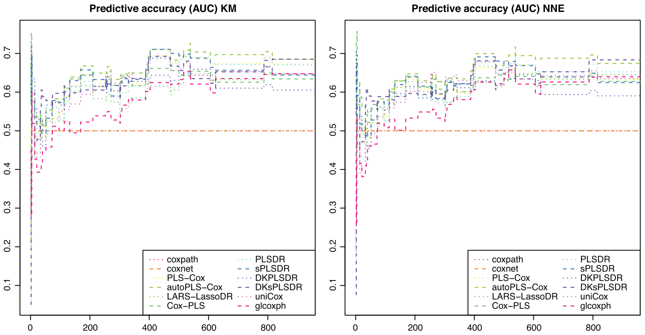

Fig. 3. Dataset 5: Survival prediction over a long period ($[0, Q_3]$)

In the simulation study, sPLSDR turned out to be the best method to recover the linear link according to the ISSw performance measure and for the three simulation schemes (Fig. 1, Linear link row panel). More generally in that setting, the models featuring components had better performance measures than the LASSO and elastic net-based ones. Similar results can be observed for the iAUCSurvROC criterion (Supplementary Fig. S8, Linear link row panel) with lesser advantage to sPLSDR and DKsPLSDR. In both cases, simulations study show that neither sPLSDR or DKsPLSDR tend to wrongly recover a link between the response and the explanatory variable when there is none (Supplementary Figs S1 and S8, No link row panel), whereas LASSO and elastic net-based methods do for the factorial simulation scheme and the iAUCSurvROC criterion (Supplementary Fig. S8, no link row panel).

Whatever the real dataset, the overall patterns of the sPLSDR and DKsPLSDR algorithms follow the general patterns of predictability of the benchmark methods, e.g. a low increase in predictability for the Metzeler dataset or a global step-wise decrease on the Romain datasets. These patterns suggest that the performances of the different methods may depend on the real but unknown data dimension. The sPLSDR or DKsPLSDR almost always rank among the 1 to 4 best methods with higher predictability, often being even 1st or 2nd, both on short and long term predictions (see Tables 2 and 3). This is particularly true in cases where a large predictability heterogeneity is to be noted among the benchmark algorithms, such as for the Garber and the Wang datasets.

Last but not least, sPLSDR and DKsPLSDR not only automatically handle missing data, the study of the robustness of these two algorithms to the amount and type of missing data

Table 2. Best methods (↘iAUC), short period ([0, Q_1]), NNE estimator

Set	1st		2nd		3rd		4th
1	**DKsPLSDR**	>	uniCox	>	glcoxph	>	coxpath
2	glcoxph	*ns*	**sPLSDR**	>	*autoPLS-Cox*	>	*PLS-Cox*
3	uniCox	>	*PLS-Cox*	*ns*	*PLSDR*	>	*autoPLS-Cox*
4	*PLSDR*	*ns*			**DKsPLSDR**	>	uniCox
	Cox-PLS	*ns*					
5	*autoPLS-Cox*	>	**sPLSDR**	*ns*	*Cox-PLS*	*ns*	**DKsPLSDR**
6	uniCox	>	**DKsPLSDR**	>	**sPLSDR**		
		>		>	*DKPLSDR*		
7	**sPLSDR**	>	*Cox-PLS*	>	*PLSDR*	*ns*	glcoxph
8	*DKPLSDR*	>			**sPLSDR**	>	**DKsPLSDR**
	coxnet	>					

Note. Wilcoxon signed-rank tests were performed to compare models according to their rankings. > stands for a significant test (5% level), *ns* for a non-significant one. We used bold fonts for the new sPLSDR and DKsPLSDR algorithms and italic for the first R implementation of existing techniques.

Table 3. Best methods (↘iAUC), long period ([0, Q_3]), KM estimator

Set	1st		2nd		3rd		4th
1	**sPLSDR**	*ns*	uniCox	*ns*	**DKsPLSDR**	>	*autoPLS-Cox*
2	*autoPLS-Cox*	>	*PLS-Cox*	*ns*	glcoxph	>	*PLSDR*
3	uniCox	>	**sPLSDR**	*ns*	**DKsPLSDR**	*ns*	*PLSDR*
4	**DKsPLSDR**	>	**sPLSDR**	*ns*			*PLSDR*
		>	*Cox-PLS*	*ns*			
5	*autoPLS-Cox*	>	**sPLSDR**	>	*PLS-Cox*	*ns*	**DKsPLSDR**
6	uniCox	>	*DKPLSDR*	>	**DKsPLSDR**	>	*PLSDR*
7	uniCox	>	**sPLSDR**				
		>	*PLSDR*				
		>	glcoxph				
8	coxnet	*ns*	*DKPLSDR*	*ns*	*autoPLS-Cox*	>	**DKsPLSDR**

Note. Wilcoxon signed-rank tests were performed to compare models according to their rankings. > stands for a significant test (5% level), *ns* for a non-significant one. We used bold fonts for the new sPLSDR and DKsPLSDR algorithms and italic for the first R implementation of existing techniques.

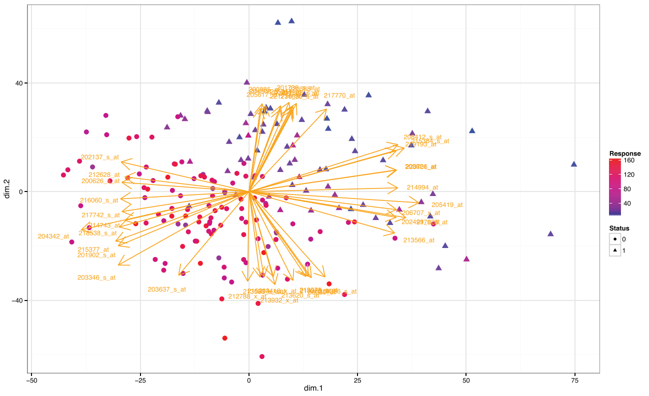

Fig. 4. Dataset 8: Biplot of individuals and most important descriptors (having any of their coordinates outside the symmetric bilateral 99.9% quantile range) on first two sPLSDR components

being beyond the scope of this article, but also provide nice data exploration tools such as biplots representation of individuals and descriptors, by projecting the dataset on the first sPLS components (see Fig. 4).

In a word, we view sPLSDR and DKsPLSDR as a useful addition to the toolbox of estimation and prediction methods for the widely used Cox's model in the high-dimensional and low-sample size settings.

ACKNOWLEDGEMENTS

The authors are grateful to the referees for their very helpful comments. They also wish to thank Caroline Chaigne and Claude Bouillon who assisted in the proofreading of the manuscript.

Funding: CNRS UMR 7501, INSERM EA3430, University of Strasbourg, L'Oréal, Recherche et Innovation and Labex IRMIA.

Conflict of interest: none declared.

REFERENCES

Akritas,M.G. (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Stat.*, **22**, 1299–1327.

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **11**, 403–503.

Antoniadis,A. *et al.* (2010) The Dantzig selector in Cox's proportional hazards model. *Scand. Stat. Theory Appl.*, **37**, 531–552.

Bair,E. *et al.* (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **101**, 119–137.

Bastien,P. (2008) Deviance residual based PLS regression for censored data in high dimensional data. *Chemom. Intell. Lab. Syst.*, **91**, 78–86.

Bastien,P. *et al.* (2005) PLS generalised linear regression. *Comput. Stat. Data Anal.*, **48**, 17–46.

Bastien,P. and Tenenhaus,M. (2001) PLS generalised linear regression, Application to the analysis of life time data. In *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*. CISIA-CERESTA Éditeur, Montreuil, pp. 131–140.

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Bennett,K.P. and Embrecht,M.J. (2003) An optimization perspective on kernel partial least squares regression. *Adv. Learn. Theory*, **190**, 227–250.

Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Bøvelstad,H.M. *et al.* (2007) Predicting survival from microarray data – a comparative study. *Bioinformatics*, **23**, 2080–2087.

Chun,H. and Keles,S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. B*, **72**, 3–25.

Collett,D. (1994) *Modelling Survival Data in Medical Research*. Chapman and Hall, London.

Cox,D.R. (1972) Regression models and life tables. *J. R. Stat. Soc. B*, **74**, 187–220.

De Jong,S. and ter Braak,C. (1994) Comments on the PLS kernel algorithm. *J. Chemom.*, **8**, 169–174.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–451.

Fan,X. *et al.* (2002) *SAS for Monte Carlo Studies: A Guide for Quantitative Researchers*. SAS publishing, Cary, NY.

Fan,J. *et al.* (2010) High-dimensional variable selection for Cox's proportional hazards model. *Borrowing Strength*, **6**, 70–86.

Fan,J. and Li,R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.*, **30**, 74–99.

Friedman,J. *et al.* (2010) A note on the group lasso and a sparse group lasso. arXiv:1001.0736.

Friedman,J.H. and Popescu,B.E. (2004) Gradient directed regularization. In: *Working Paper*, Department of Statistics, Stanford University.

Fu,W.J. (1998) Penalized regression: the bridge versus the LASSO. *J. Comput. Graph. Stat.*, **7**, 397–416.

Garber,M. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.

Garthwaite,P.H. (1994) An interpretation of partial least squares. *J. Am. Stat. Assoc.*, **89**, 122–127.

Goeman,J.J. (2010) L_1 Penalized estimation in the cox proportional hazards model. *Biometric. J.*, **52**, 70–84.

Gonen,M. and Heller,G. (2005) Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**, 1809–2005.

Gui,J. and Li,H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with application to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Harrell,F.E. *et al.* (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. *Stat. Med.*, **15**, 361–387.

Heagerty,P.J. *et al.* (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Heagerty,P.J. and Zheng,Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics*, **61**, 92–105.

Jolliffe,I. *et al.* (2003) A modified principal component technique based on the lasso. *J. Comput. Graph. Stat.*, **12**, 531–547.

Kaiser,H.F. and Dickman,K. (1962) Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, **27**, 179–182.

Kim,J. *et al.* (2008) A gradient-based optimization algorithm for lasso. *J. Comput. Graph. Stat.*, **17**, 994–1009.

Krämer,N. (2007) An overview on the shrinkage properties of partial least squares regression. *Comput. Stat.*, **22**, 249–273.

Lambert-Lacroix,S. and Letué,F. (2011) Partial least squares and cox model with application to gene expression. *Technical report*. http://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseVI/publications_2011/TR/Lambert-LacroixLetueIAP.pdf.

Langfelder,P. *et al.* (2013) When is hub gene selection better than standard meta-analysis? *PloS One*, **8**, e61505.

Li,H. and Gui,J. (2004) Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20**, 208–215.

Li,L. (2006) Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, **22**, 466–471.

Lindgren,F. *et al.* (1993) The kernel algorithm for PLS. *J. Chemom.*, **7**, 45–59.

Magidson,J. and Wassmann,K. (2010) The role of proxy genes in predictive models: an application to early detection of prostate cancer. In: *Joint Statistical Meetings Proceedings*. American Statistical Association, Vancouver, pp. 2739–2753.

Marx,B.D. (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38**, 374–381.

Metzeler,K.H. *et al.* (2008) An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, **112**, 4193–4201.

Nguyen,D.V. and Rocke,D. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Park,M.Y. and Hastie,T. (2007) L_1 regularization path algorithm for generalized linear models. *J. R. Stat. Soc. B*, **69**, 659–677.

Rännar,S. *et al.* (1994) A PLS kernel algorithm for data sets with many variables and fewer objects. Part I: theory and algorithm. *J. Chemom.*, **8**, 111–125.

Romain,B. *et al.* (2010) Allelotyping identification of genomic alterations in rectal chromosomally unstable tumors without preoperative treatment. *BMC Cancer*, **10**, 561.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Rosipal,R. and Trejo,L.J. (2001) Kernel partial least squares regression in reproducing kernel hilbert space. *J. Mach. Learn. Res.*, **2**, 97–123.

Schmid,M. *et al.* (2011) A robust alternative to the Schemper-Henderson estimator of prediction error. *Biometrics*, **67**, 524–535.

Segal,M.R. (2006) Microarray gene expression data with linked survival phenotypes: diffuse large-Bcell lymphoma revisited. *Biostatistics*, **7**, 268–285.

Simon,N. *et al.* (2011) Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, **39**, 1–13.

Simon,N. *et al.* (2013) A sparse-group Lasso. *J. Comput. Graph. Stat.*, **22**, 231–245.

Sohn,I. *et al.* (2009) Gradient lasso for cox proportional hazards model. *Bioinformatics*, **25**, 1775–1781.

Tenenhaus,A. *et al.* (2007) Kernel logistic PLS: a tool for supervised nonlinear dimensionality reduction and binary classification. *Comput. Stat. Data Anal.*, **51**, 4083–4100.

Tenenhaus,M. (1998) *La régression PLS*. Technip, Paris.

Tenenhaus,M. (1999) La regression logistique PLS. In *Proceedings of the 32èmes journées de Statistique de la Société française de Statistique*. Fès, pp. 721–723.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tibshirani,R. (2009) Univariate shrinkage in the cox model for high dimensional data. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–18.

van Wieringen,N. *et al.* (2009) Survival prediction using gene expression data: a review and comparison. *Comput. Stat. Data Anal.*, **53**, 1590–1603.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wang,S. *et al.* (2009) Hierarchically penalized Cox regression with grouped variables. *Biometrika*, **96**, 307–322.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. Krishnaiah,P.R. (ed.) (1982) *Multivariate Analysis*. Academic Press, New York, NY, pp. 391–420.

Wold,S. *et al.* (1983) The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe,A. and Kåstrøm,B. (eds) *Proceeding Conference Matrix Pencils, March 1982, Lecture Notes in Mathematics*. Springer Verlag, Heidelberg, pp. 286–293.

Xu,R. and O'Quigley,J. (1999) A R^2 type measure of dependence for proportional hazards models. *J. Nonparametr. Stat.*, **12**, 83–107.

Zhang,H.H. and Lu,W. (2007) Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, **67**, 301–320.

Zou,H. *et al.* (2006) Sparse principal component analysis. *J. Comput. Graph. Stat.*, **15**, 265–286.