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, Aulnay, France, pbastien@rd.loreal.com
[2] IRMA, CNRS UMR 7501, Labex IRMIA, Université de Strasbourg, France, fbertran@math.unistra.fr., mmaumy@math.unistra.fr
[3] EA3430, INSERM. Laboratoire de Biostatistique, Faculté de Médecine de Strasbourg, France, nmeyer@unistra.fr

Keywords: PLS, Cox, Sparse, Deviance, Kernel.

1 Introduction

There has been a vast literature in the last decade devolved to relating gene expression profiles to subject's survival or to time to cancer recurrence. Biomarker discovery from high dimensional data, such as transcriptomic or SNP profiles, is a major challenge in order to allow more precise diagnosis. The proportional hazard regression model suggested by Cox, 1972 [1], to study the relationship between the time to event and a set of covariates in the presence of censoring is the model most commonly used for the analysis of survival data. However, like multivariate regression, it supposes that there are more observations than variables, complete data, and variables not strongly correlated between them. 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. In order to deal with both collinearity and variable selection issues, many methods based on Lasso penalized Cox proportional hazard have been proposed since the seminal paper of Tibshirani, 1997[2] : Fan and Li (2002), Gui and Li (2005), Segal (2006), Park and Hastie (2007), Zhang & Lu (2007), Zou (2008), Sohn (2009),Goemann (2009), Tibshirani (2009), Fan et al (2010), Simon et al (2011).

Regularization could also be performed using dimension reduction as is the case with PLS regression. PLS regressions has already been extended to Cox regression (Bastien & Tenenhaus (2001) [3], Nguyen and Rocke (2002), Li and Gui (2004)). Recently, Chun & Keles (2010) [4] provide both empirical and theoretical results that the performance of PLS regression is ultimately affected by the large number of predictors. In particular, existence of higher number of irrelevant variables leads to inconsistency of coefficient estimates in linear regression setting. Chun & Keles proposed sparse PLS regression which promotes variables selection within the course of PLS dimension reduction. Moreover, they demonstrated that for univariate PLS, the first direction vector of their sparse PLS algorithm is easily obtained by soft thresholding of the original PLS direction vector.

In 2006, Segal showed [5] that the expression to be minimized in the Cox-Lasso procedure of Tibshirani can be approximated, in a first order Taylor-series approximation sense, by the deviance residual sum of squares, a normalized transform of the martingale residuals. Following Segal, we proposed in 2008 [6] an alternative to the PLS-Cox model in high dimensional settings by using deviance residual based PLS regression.

Using results from Segal and Chun & Keles, we propose [7], two original algorithms named sPLSDR and its nonlinear kernel counterpart DKsPLSDR, by using sparse PLS regression (sPLS) based on deviance residuals. We compared their predicting performance with state of the art algorithms based on both simulated and reference benchmark datasets.

2 Material and methods

The SPLSDR algorithm involves the following steps:

1) Cox model without covariates in order to compute the null deviance residuals.

2) Computation of the sPLS components by using the sPLS regression with the null deviance residuals as outcome.

a) Set $\hat{\beta}^{PLS} = 0, \mathrm{A} = \{\ \}, k = 1, y_1 = d$

b) while ($\mathrm{k} \le K$)

(1) $w = \left(|z| - \lambda/2\right)_+ sign(z)$ where $z = X'y_1 / \|X'y_1\|$

(2) Update A $as \{i : \hat{w}_i \neq 0\} \cup \{i : \hat{\beta}_i^{PLS} \neq 0\}$

(3) Fit PLS with X_A 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 (cross-validation) sPLSDR components.

3 Results and discussion

The two algorithms sPLSDR and DKsPLSDR compare favorably with other methods in their computational time, prediction and selectivity, based on results from benchmarks and simulated datasets. They not only automatically handle missing data using the NIPALS algorithm, but also provide nice data exploration tools such as biplot representations of individuals and descriptors, by projecting the dataset on the first sPLS components.

As a result, 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.

The R-package plsRcox is available on the CRAN and is maintained by Frédéric Bertrand.

4 References

[1] Cox, D.R., Regression models and life tables. *JRSSB*, **74**, 187–220, 1972.

[2] Tibshirani R., The lasso method for variable selection in the Cox model, *Statistics in Medicine* 16 (1997) 385–395..

[3] Bastien P., Esposito Vinzi V., and Tenenhaus M., PLS Generalised Linear Regression, *Computational Statistics & Data Analysis*, 48, 17-46, 2005.

[4] Chun, H., and Keles, S., Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *JRSSB*, **72**(1), 3–25, 2010.

[5] Segal M.R., Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited, *Biostatistics* 7 268–285, 2006.

[6] Bastien P., Deviance residuals based PLS regression for censored data in high dimensional setting, *Chemometrics and Intelligent Laboratory Systems* 91, 78-86, 2008

[7] Bastien P., Bertrand F., Meyer N. ,Maumy-Bertrand M., Deviance residuals based sparse PLS and sparse kernel PLS regression for censored data, Bioinformatics (2015) 31 (3): 397-404..