PLS-Cox model : Application to gene expression.

Philippe Bastien

L'Oréal Recherche, 1 avenue Eugène Schueller – BP 22 – 93601 Aulnay sous Bois Cedex
pbastien@recherche.loreal.com
Key Words : Cox model, PLS Regression, Gene profiling
COMPSTAT 2004 section : Partial Least Squares

ABSTRACT

With advances in high-density DNA microarray technology, gene expression profiling is extensively used to discover new markers and new therapeutic targets. This technique supposes to take into account the expression of thousands of genes with respect to a limited number of patients. To predict survival probability on the basis of gene expression signatures can become a very useful diagnostic tool. In the context of highly multidimensional data the classical Cox model does not work. The PLS-Cox model by operating a dimension reduction of the gene expression space directed towards the explanation of the risk function appears particularly useful. It allows the determination of signatures of genomic expressions associated with survival, to predict the survival probability from these profiles, and reduce inter individual variability by changing the level of adjustment from a phenotypical level to a genotypical level.

I.INTRODUCTION

The proportional hazard regression model suggested by Cox in 1972 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 models, it supposes that there are more observations than variables, complete data, and variables not strongly correlated between them. These constraints are often crippling in practice. In particular the analysis of transcriptomic data supposes to take into account the expression of thousands of genes compared to only a limited number of patients. The solution suggested is to initially operate a dimension reduction of the space of genes directed towards the explanation of the risk function. One then builds a Cox model on the PLS components.

Alizadeh et al. (2000) identified from the expression of genes of 40 subjects suffering from diffuse large B-cell lymphomas (DLBCL) two subgroups, each characterized by a distinct gene expression signature. These were associated with very different clinical prognoses. Using information on patients survival allows the determination of genotypic signatures linked to the risk function. Survival probabilities have then been carried out from these expression

profiles. We show that these genotypic signatures bring additional informations to an index of existing clinical risk.

II METHODS

The suggested method associate PLS regression (Wold, Martens, Wold 1983) with the Cox model. It has already been used on epidemiological data (Bastien, Tenenhaus, 2001). Its specificity is that it takes into account the censoring information in the construction of PLS components.

PLS-Cox algorithm

Let $X_0=\{x_1,...,x_p\}$ a matrix whose columns are gene expression (log ratio). One seeks successively m orthogonal PLS components T_h which are linear combinations of the x_j. In particular the research of the h-th.PLS component T_h is carried out according to the following steps:

Step 1 : For j=1 to p, calculate the coefficients of regression a_{hj} of x_j in the Cox model with covariates T_1, T_2,..., T_{h-1} and x_j.

Step 2: normalize the column vector a_h formed by a_{hj} : $w_h = a_h/||a_h||$

Step 3: calculate the residual X_{h-1} of the linear regression of X_0 on T_1...,T_{h-1}

Step 4: calculate the component $T_h = X_{h-1}w_h / w_h'w_h$.

Step 5: express the component T_h according to X_0 : $T_h= X_0 w_h^*$

The prediction of the risk function h(t) is then carried out in a natural way with the Cox model adjusted on PLS components. The regression equation can also be written according to the original data with the coefficients confidence intervals estimated by bootstrap resampling.

Cross-validation

The number k of PLS components T_h was chosen by cross-validation. Each patient's score was estimated using a training data set of N-1 samples (leave-one-out CV).

Let i be the subscript for sample i and -i the subscript when sample i is leaved out. The score for patient i on h-th PLS component is defined as :

$$t_{h,i}=x_{h-1,i}w_{h,-i}=\left(x_{0,i}-\sum_{j=1}^{h-1}t_{j,i}p'_{j,-i}\right)w_{h,-i}$$

with : $x_{h-1,i}$ the i^{th} row of the residual matrix X_{h-1}

$w_{h,-i}$ the weights based on $X_{h-1,-i}$

$p'_{j,-i}$, the loadings, defined as the coefficients of $T_{j,-i}$ in the regression of $X_{0,-i}$ on $T_{1,-i},\ldots,T_{j,-i}$, the j first PLS components carried out on $X_{0,-i}$

PLS-Cox and PLS-GR

The Cox-PLS algorithm uses the principles of the NIPALS algorithm (Wold 1966) and can also function in the presence of missing data. The PLS-Cox model is a particular case of PLS generalized linear regression (Bastien, Esposito Vinzi, Tenenhaus, 2004).

Estimation of the survivor function

During the prediction phase, a proportional hazard model is fitted with the k PLS scores T_1, ..T_k as covariates. Let $S_o(t) = \exp\left[-\int_0^t h_0(u)du\right]$ the baseline unspecified survivor function, $T'_i=(t_{i1},\ldots,t_{ik})$, and $\beta'=(\beta_1,\ldots,\beta_k)$. The survivor function given the scores T_i is : $S(t,T_i)=S_0(t)^{\exp(T_i\beta)}$. The calculation of the non-parametric maximum likelihood of $S_0(t)$ (Kalbfleich and Prentice, 1973) is based on the product limit estimate with similar argument to that used in obtaining kaplan-Meier estimate.

Let $t_{(1)},\ldots, t_{(l)}$ be the distinct failures times, the likelihood function is maximized by taking $S_0(t) = S_0(t_{(j)}+0)$ for $t_{(j)} < t \leq t_{(j+1)}$ and allowing probability mass to fall only at the observed failure time $t_{(j)}$. This leads to the consideration of a discrete model with hazard contribution $1-\alpha_j$ at $t_{(j)}$.

let $\hat{S}_0(t) = \prod_{j/t(j)<t} \hat{\alpha}_j$ the survival probability at time t

$\hat{S}_i(t) = \prod_{j/t(j)<t} \hat{\alpha}_j^{\exp T_i\,\beta}$ the survival probability at time t for a patient with covariates T_i

The maximum likelihood estimate $\hat{\alpha}_j$ of α_j is obtained numerically from : $\sum_{k\in Fj} \frac{\hat{u}_k}{1-\hat{\alpha}_j^{\hat{u}_k}} = \sum_{l\in R(t_{(j)})} \hat{u}_l$

with $\hat{u}_k=\exp(T_k'\hat{\beta})$, F_j the set of individuals failing at time $t_{(j)}$ and $R(t_{(j)})$ the risk set at time $t_{(j)}$. In case where there are no ties then the set F_j contains only one individual and the solution to the above equation can be solved analytically and is given by $\hat{\alpha}_j=[1-(\hat{u}_j/\sum_{l\in R(t_{(j)})}\hat{u}_l)]^{\hat{u}_j^{-1}}$. One finds the Kaplan-Meier estimator when $T_i = 0$ for all the individuals : $\hat{S}(t) = \prod_{j/t(j)<t} \frac{(n_j-d_j)}{n_j}$

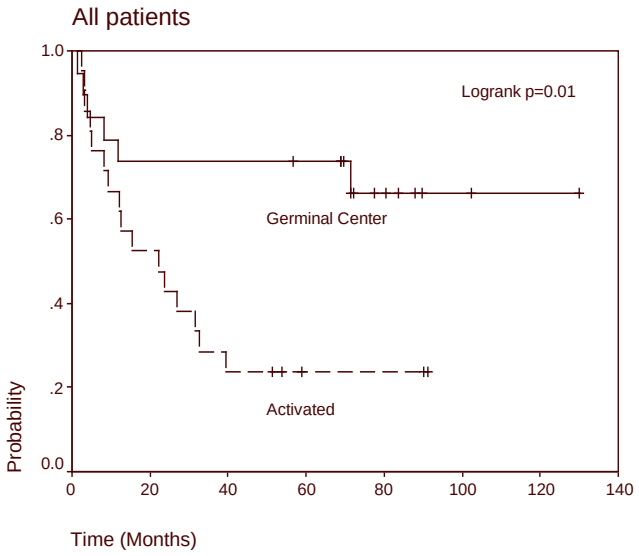
III APPLICATION

The data set from Alizadeh and al. consists of gene expression level from cDNA experiments involving three prevalent adult lymphoid malignancies : Diffuse large B-cell lymphoma (DLBCL), B-Cell chronic Lymphocytic Leukemia (BCLL), and Follicular Lymphoma (FL). Data are available on the study supplement web site (http://llmpp.nih.gov/lymphoma/data.shtml) .

cDNA targets were prepared from experimental mRNA samples and were labelled with Cy5-dye during reverse transcription. A reference cDNA sample was prepared from a combination of nine different lymphoma cell lines and was labelled with Cy3-dye. Cy-labelled experimental and reference cDNAs were mixed and hybridised onto the microarray The standardized intensity ratio of fluorescence was then quantified for each gene. It reflects the relative abundance of the gene in each experimental sample of mRNA compared to the reference sample.

By using clustering analysis, Alizadeh and al. identified two DLBCL sub-groups with different transcriptomic profiles. They correspond to distinct levels of lymphocytes B differentiation: Germinal Center B-like (19 patients) and Activated B-like (21 patients).
In complement to the transcriptomic data, the duration of patients survival was also collected. Among the 40 patients 22 events (death) were observed and the 18 remaining survival durations being censured. Patients with a DLBCL of the Germinal center B-like have, on average, a significantly better survival than those with Activated B-like type as shown on figure 1. The molecular classification of the tumours on the basis of their genetic expression profile thus allows to highlight sub-types of cancer non identified.

Figure 1 : Kaplan-Meier survival curves estimates by molecular sub-groups

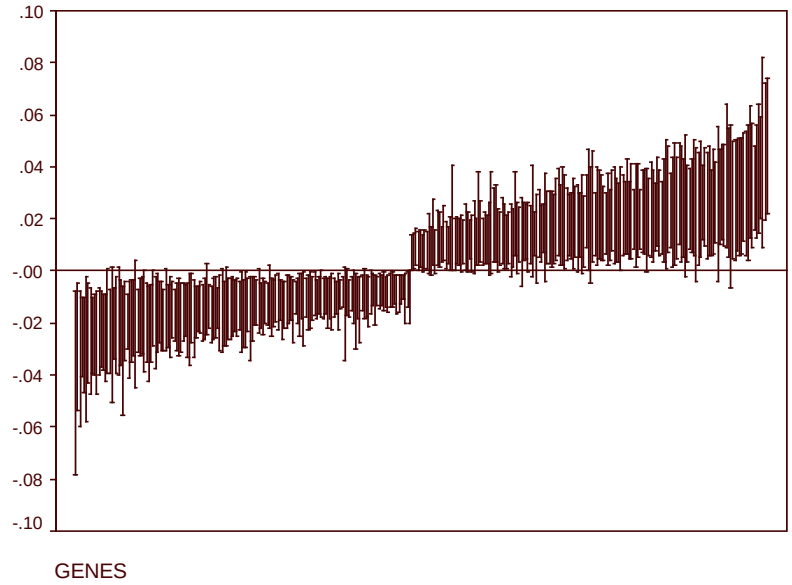


IV RESULTS

One Thousand and height hundred genes were selected from over more than 13000 to having different expressions to the two molecular types (ttest, $p< 0.05$). We retained two PLS components by cross-validation. Once PLS-Cox model has been estimated, the significance of genes coefficients could be ascertain in a non parametric framework by means of a bootstrap procedure. Bootstrap confidence intervals were computed based on the 2.5 and 97.5 percentiles of the bootstrap empirical distribution (balanced bootstrap, B=500).

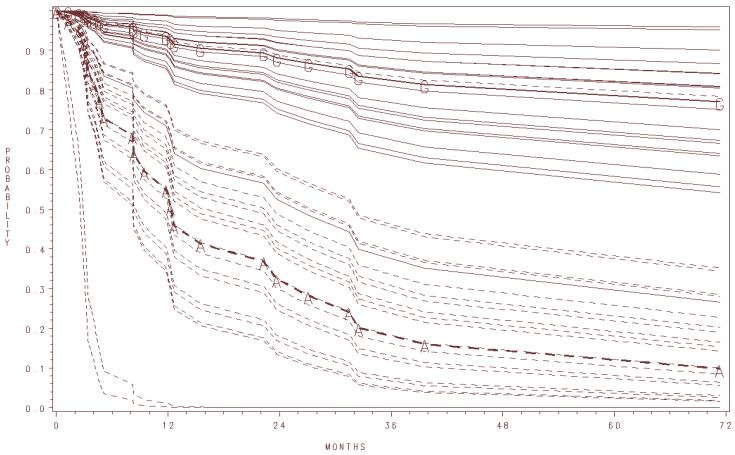
Figure 2 presents coefficients confidence intervals of the PLS-Cox model on two components expressed according to their original data (log ratio). The coefficients were sorted by ascending values. In order to simplify PLS components, only genes having a significant contribution at the 5% threshold were taken into account. It explains the clear separation on both sides of the ordinate axis.

Figure 2 : 95% bootstrap confidence intervals for genes coefficients



The following graph (figure 3) presents the individual distributions for the patients of the Activated B-like type (dotted line) and for those of Germinal center B-like type (continuous line). The letters represent the average distributions by molecular type. The distributions were estimated by cross-validation with two PLS components.

Figure 3 : Cross-validated survival curves

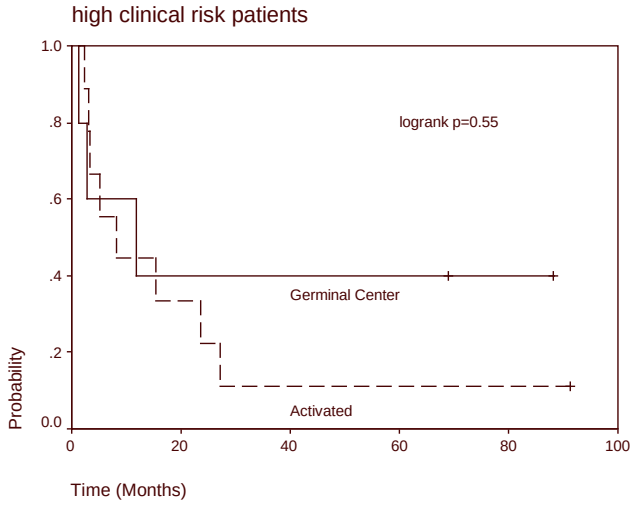


Based on the log-ratio of gene expressions for the mean levels of PLS components, the survival curves demonstrated more marked prognoses between the two molecular types in comparison to the survival estimation using Kaplan-Meier. The genotypic signatures of the two molecular types appear well associated with the different prognoses, with very minor overlaps.

International Prognostic indicator (IPI)

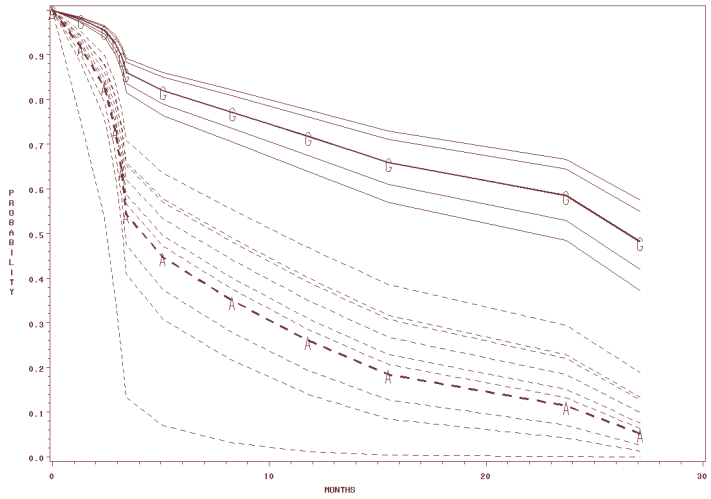
A clinical index scored from 0 to 5 is used to define sub-groups of patients suffering from DLBCL. The subjects of the group with the lowest scores IPI (0-2) have a better prognostic than those having highest scores (3-5). Alizadeh et al showed that in the group with the lowest risk factors, the patients presenting a profile of genetic expression of Germinal center B-like type had a significantly better survival (Logrank, $p<0.05$) than those of Activated B-like type. They did not observe a similar effect in the higher risk factors group (Logrank, $p=0.55$) as illustrated in figure 4.

Figure 4 : Kaplan-Meier survival curves for the high clinical risk patients



The PLS-Cox model on the higher risk factors group, taking into account the transcriptomic information is more selective and allows the differentiation of the two molecular types. Figure 5 shows the individual distributions of survival estimated by cross-validation.

Figure 5 : Cross-validated survival curves for the high clinical risk patients



The gene expression signature makes it possible to differentiate the two molecular types. More precise clinical diagnosis procedures could therefore be developed.

V DISCUSSION

With advances in high-density cDNA microarray technology, gene expression profiling is extensively use to discover new markers and new therapeutic targets. This technique supposes to take into account the expression of thousands of genes with respect to only a limited number of patients. To predict survival probability on the basis of gene expression signatures can become a very useful diagnostic tool. In the context of highly multidimensional data the classical Cox model does not work.

Recently Nguyen and Rocke (2002) illustrated using the example of Alizadeh et al. the use of PLS components as covariates to predict the probabilities of survival in a Cox model. However their model was not completely satisfactory, since it did not take into account the censoring information in the estimation of PLS components, thus inducing a potential bias in their estimates.

The PLS-Cox model described above shows major improvement with respect to the method proposed by Nguyen and Rocke. It takes into account the censoring information in the estimation of PLS components. In case of missing data, PLS components are computed in

accordance with the NIPALS algorithm. Moreover statistical significance of gene coefficients is ascertain using bootstrap validation procedure.

The PLS-Cox model by operating a dimension reduction of the genes expression space directed towards the explanation of the risk function appears particularly useful. It allows the determination of signatures of genomic expressions associated with survival, to predict the survival probability from these profiles, and reduce inter individual variability by changing the level of adjustment from a phenotypical level to a genotypical level. In order to assess the efficacy of new drugs, study design will benefit from a better characterisation of patient groups made possible by genomic expression profiling.

VI REFERENCES

[1] Allison, Paul D. (1995) : *Survival Analysis Using the SAS System* : A practical guide, SAS Inc, Cary, NC.

[2] Alizadeh,A.A. et al.(2000). *Distinct types of diffuse large B-cell lymphoma identified by gene expression profile*. Nature,403, 503-511.

[3] Bastien P., Tenenhaus M. (2001) : PLS generalized linear regression. Application to the analysis of life time data. In *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*, Esposito Vinzi V., Lauro C., Morineau A. & Tenenhaus M. (Eds). CISIA-CERESTA Editeur, Paris, p. 131-140.

[4] Bastien P.,Esposito Vinzi V., Tenenhaus M (2004)., PLS generalized linear regression, Computational Statistics & Data Analysis, to appear

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

[6] Kalbfleich J.D. and Prentice R.L. (1973) *Marginal Likelihoods based on Cox's regression and life model*. Biometrika, 60, 267-278.

[7] Nguyen D.V. and Rocke D. (2001) *Partial least squares proportional hazard regression for application to DNA microarray survival data*, Bioinformatics, 18, 1625-1632.

[8] Tenenhaus M. (1998) : *La régression PLS*. Technip, Paris

[9] Wold S., Martens & Wold H. (1983) : The multivariate calibration problem in chemistry solved by the PLS method. In *Proc. Conf. Matrix Pencils*, Ruhe A. & Kåstrøm B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, p. 286-293.

[10] Wold H.,(1966) : *Estimation of principal components and related models by iterative least squares*, in Multivariate Analysis, Krishnaiah P.R.(Ed.), Academic Press, New York, pp. 391-420.