Using regularization patterns from penalized regression in microarray experiments

Giulia Tonini[a,b], Michela Baccini[a,b], Duccio Cavalieri[c], Enrico Mini[c], Piero Dolara[c], Annibale Biggeri[a,b]

[a] Dept. of Statistics "G. Parenti", University of Florence, viale Morgagni 59, 50134 - Florence (Italy)
[b] Biostatistics Unit, ISPO, via Cosimo il Vecchio 2, 50139 - Florence (Italy)
[c] Dept. of Pharmacology, University of Florence, viale Pieraccini 6, 50139 - Florence (Italy)

Corresponding Author:
Giulia Tonini
g.tonini @ispo.toscana.it

Summary
Microarray experiments have been used to investigate the relationship between gene expression and survival in cancer patients. Many methods have been developed, but most of them do not take into account other prognostic variables and the interplay among genes. A solution consists in using penalized regression models for censored survival data. We propose a single graphical approach to complement penalized regression analysis. In this paper, we illustrate the methodology using a penalized Cox regression approach on two different microarray data sets (colon and lung cancer). A small simulation study completes the paper. On both data sets, we applied a L2 penalized Cox Regression, after having pre-selected the most relevant sets of gene expression data according to a generalized logrank test. The patterns of estimated gene expression coefficients were explored varying the penalty parameter.
We compared these results with those obtained while using a L1 penalty and found that the two approaches gave consistent, but not identical, results. The simulation study confirms the results for three different correlation scenarios.
Theoretical considerations indicate that the L2 penalized regression is a more appropriate approach in this context. We propose to consider the entire regularization pattern varying the penalty parameter as a graphical tool in a sensitivity analysis.

Introduction

Gene expression experiments are potentially useful to identify subgroups of patients with good/bad prognosis. The advantage of functional genomics assessment is substantial when classical prognostic markers have limited value.
For example in the seminal article by Alizadeh et al. [1] the molecular classification of tumors on the basis of gene expression identified previously undetected and clinically significant subtypes of cancer. Many microarray studies focus on survival time of patients as the primary clinical outcome. The prediction of survival is the main issue of many published papers. A review can be found in [27].
A popular approach to screen for genes candidates as prognostic markers consists in first classifying patients on the basis of gene expressions and second in evaluating if and to which extent the identified subgroups experience differential survival [10]. This approach, while widely used, is inefficient because there could be differences in genomic expressions identified at the first step that do not relate to prognosis and viceversa. The alternative practice,with censored life histories, of comparing dead/alive subjects on the basis of gene expression is incorrect.
A more efficient approach consists in directly relating the gene expression profile to survival [10]. For this reason, among others, the Significance Analysis of Microarrays (SAM) [26] has been generalized for working with censored survival data (survival option in the samr function of R software). These procedures are usually marginal: they investigate the effect of gene expression without taking into account other potential prognostic variables (i.e. age, gender, stage of disease) and they assess the effect of each gene separately.

Estimating the net contribution of gene expression in predicting survival, once patient and tumor characteristics are accounted for, is a scientifically sound goal [12]. Moreover, since good/bad prognosis may result from the interplay between many genes, methods that simultaneously use data from many genes are expected to have better performance in explaining risk than methods that investigate each gene separately. In microarray experiments, the number of genes is much larger than the number of patients. Consequently,because of the singularity of the design matrix, including a large number of genes in the same model is a major problem. A solution for this problem is to use methods based on penalized regression [24]. With penalized regression, the conditional effect of each single gene expression, given the rest of the genomic information, can be evaluated by the introduction of a constraint on gene expression coefficients.

The constraint is defined in order to reduce the effective number of parameters included in the model (see for example [14]).

Several constraint definitions can be used. Ridge regression and Lasso regression are examples of penalized regressions that shrink coefficients towards zero by L2 and L1 constraint, respectively. The Lasso penalty is defined as

$$pen(\beta) = \sum_{j=1}^{p} |\beta_j| \qquad [1]$$

i.e. the $L1$ norm of the regression coefficient, while the Ridge penalty takes into consideration the $L2$ norm

$$pen(\beta) = \sum_{j=1}^{p} |\beta_j|^2 \qquad [2]$$

Ridge regression was first suggested by Hoerl [16] to overcome situation in which correlations between the predictors give rise to unstable parameter estimates. L1 penalized regression has been reviewed by Lokhorst [17]. An important property of the L1 penalty is that it can generate exact zero estimates of the coefficients. This means that it can be viewed as an automatic variable selection method, where some of the variables are eliminated from the model as the penalization becomes stronger [18].

Recently penalized regression was extended to censored survival data [25].

Several modified approaches to penalized Cox regression models have been proposed [29][7]. Usually penalized Cox regression is combined with cross validation methods in order to select the best degree of constraint [5].

In this paper we use a Ridge regression approach to the analysis of censored survival data from microarray experiments.

Our aim is to assess the net effect of a given gene while adjusting for other genes. We propose to avoid selection of the best degree of constraint in penalized regression, but to consider the behavior of the estimated coefficients varying the degree of shrinkage. Similar approaches can be found in literature [16].

We provide an example of this approach by using two real data sets. The first data set comes from a study on colorectal cancer [6] and the second from a study on lung cancer [4]. In both applications, after having adjusted for other relevant prognostic factors and after taking into consideration the whole genomic information, we considered the net contribution of each gene on survival. In doing this, for reasons of simplicity, we preliminarily selected a subset of candidate genes, followed by specification and fitting of a L2 penalized Cox regression model. For comparison, we applied on the same data sets a L1 penalized regression. Comparison between L1 and L2 penalized regressions in analysis of censored survival data can be found also in [5]. We then carry out simulations to further illustrate the performance of the methods.

Data

ITT Colon Cancer Data

The first dataset was relative to a study sponsored by ITT (Istituto Toscano Tumori), which was aimed at discovering potential markers of prognosis in colorectal cancer [6]. This study was restricted to Duke's stage C and D, G2 grade, adenocarcinoma histotype. All patients had surgery and fluoropyrimidine-based chemotherapy. Primary tumors were obtained immediately after resection. Gene expression profiles were obtained from 19 patients

using cDNA microarrays. The dataset considered in this analysis was comprised of 2587 genes overall. All the patients were followed up on until their death or until end date of the study. The study started on June 1, 1994, ended on March 30, 2006. We observed 13 deaths. The median value of survival obtained through the Kaplan-Meier estimator was 38.63 months. Main clinical markers were Duke's stage, C1/C2 and D, location of tumor (rectum, junction, sigma and transverse colon), age at surgery and gender. According to the Duke's classification, 7 patients were classified as stage C1, 5 as stage C2 and 7 as stage D. In 11 patients the tumor was localized in the transverse, sigma or right colon, while in 8 patients it was localized in the rectum. Patient age at time of surgery was between 46 and 71 years old.

Bhattacharjee's Lung Cancer Data

The second dataset was related to lung cancer patients [4]. The data consists of gene expressions of 12600 genes for 125 patients. The patients were classified according to the progression of the disease. 61 patients were classified as stage I; 36 as stage II; 18 as stage III; 10 as stage IV. For each of the 125 patients, the survival time as well as the censoring status was available. There were 63 failures. The median value survival obtained through the Kaplan-Meier estimator was 37.6 months. Information about patients's age was also available.

Methods

The statistical analysis was based on a two-steps procedure [20]. In the first step of the analysis, once we have accounted for the effect of other relevant prognostic factors, we ranked genes according to their ability to predict survival. Then, in the second step of the analysis, the list of the K top genes was included in the penalized regression models. This step is necessary to adapt the penalized Cox regression to microarray data. The typical microarray dataset contains thousands of genes, but the algorithm tends to become slow when the number of covariates is much higher than the number of samples. Applying the method to a rank-ordered list of genes is a common solution, [3].

First step: Preliminary Selection

The preliminary selection was done specifying a Cox regression model for each gene separately. It aims to select a subset of candidate genes for the subsequent penalized regression analysis. We started from a "core" model which did not include gene expression, but took into account for other prognostic variables ($z1$, $z2$, ..., zm):

$$h(t; z_1, z_2, ..., z_m) = h_0(t) \exp(\sum_{i=1}^{m} \gamma_i z_i); \quad [3]$$

where $h(t; z1, z2, ..., z$m$)$ denotes the hazard function, given the values of the covariates ($z1$, $z2$, ..., zm), ($\gamma 1$, $\gamma 2$, ..., γm), are unknown parameters and h0(t) is the baseline hazard function.

For the first data set, the core model included as covariates sex, age at time of surgery (≤ 65,> 65), Dukes stage of the tumor (C1, C2, D) and tumor location (not rectum, rectum). For the second data set, stage of the tumor (I, II, III, IV) and patient age (< 50,$50 - 70$,> 70) were considered. Then we extended the "core" model adding a linear term for the relative expression values of each gene. For the gth gene, the extended model was the following:

$$h(t; z_1, z_2, ..., z_m) = h_0(t) \exp(\sum_{i=1}^{m} \gamma_i z_i); \quad [4]$$

where g_k is the expression value of the kth gene and β_k is the unknown gene specific regression coefficient.

Gene ranking was based on the Rao's generalized score test statistic for the coefficients β_k. Let $U(\gamma 1, \gamma 2, ..., \gamma$m, $\beta_k)$ be the partial score vector and $J(\gamma 1, \gamma 2, ..., \gamma$m, $\beta_k)$ be the $(m + 1) \times (m + 1)$ observed information matrix for the model (4). The generalized score test statistics for β_k can be obtained from the following equation:

$$X^2 = U(\hat{\gamma_1}, \ldots, \hat{\gamma_m}, 0)^T [J(\hat{\gamma_1}, \ldots, \hat{\gamma_m}, 0)]^{-1} U(\hat{\gamma_1}, \ldots, \hat{\gamma_m}, 0), \quad [5]$$

where $(\hat{\gamma} 1, \hat{\gamma} 2, ..., \hat{\gamma}$m) m) is the vector of the Partial Maximum Likelihood estimates of the "core" model (3) coefficients.

The advantage of ranking according to the score test statistics is that having the "core" model (3)

fitted the X^2 statistics can be obtained directly without fitting the extended model (4). This implies that with a strong gain in terms of computational burden the iterative procedure for maximization of the Partial Likelihood of the extended model is not required. This is particularly true here, due to the large number of genes to be included, because we were considering many extensions of the core model.
More details on the score test calculation are presented in the appendix.
Ad hoc codes were written in *R*-language for calculating the generalized score test statistics and for ranking genes (available on request).

Second step: Penalized Regression

In the second step of the analysis, the subset of K candidate genes selected according to the score test statistics were included in a penalized Cox regression model. Let $(g_1, g_2, \ldots, g_K)$ be the vector of relative expression values of the K selected genes. These were included simultaneously in the Cox model:

$$h(t; z_1, z_2, ..., z_m, g_1, g_2, \ldots, g_K) = \\ = h_0(t) \exp(\sum_{i=1}^{m} \gamma_i z_i + \sum_{k=1}^{K} \phi_k g_k) \quad [6]$$

Then a constraint was specified on the gene expression coefficients $(\phi_1, \phi_2, \ldots, \phi_k)$. We considered a quadratic constraint following a Ridge regression:

$$\sum_{k=1}^{K} \phi_k^2 < s. \quad [7]$$

The introduction of the constraint reduces the effective number of parameters to be estimated. In our case, this allowed the estimation of a model where the number of parameters is higher than the number of observations. The consequence of such constrained estimation is that all gene expression coefficients $(\phi_1, \ldots, \phi_K)$ are shrunk toward zero, while the coefficients relative to the other prognostic variables are left unconstrained.
The model (6), under the constraint (7), can be estimated maximizing the penalized partial log-likelihood:

$$\ell_\theta(\gamma_1, \gamma_2, \ldots, \gamma_m, \phi_1, \phi_2, \ldots, \phi_K) = \\ \ell(\gamma_1, \gamma_2, \ldots, \gamma_m, \phi_1, \phi_2, \ldots, \phi_K) - \frac{\theta}{2} \sum_{k=1}^{K} \phi_k^2, \quad [8]$$

for some number $\theta \geq 0$. The function $\ell(..) = log(L(..))$ is the Partial Log-Likelihood where in general

$$L(\alpha) = \Pi_{Y_i uncens} \frac{exp(X_i \alpha)}{\sum_{Y_j \geq Y_i} exp(X_j \alpha)} \quad [9]$$

θ in (8) is usually referred to as the smoothing, regularization or penalty parameter. θ is a function of s and controls the amount of shrinkage of the constrained parameters. Large values of θ corresponds to a strong constraint, while small values of θ give less smoothed coefficients. θ has a very interesting Bayesian interpretation: it represents the prior expectation on the magnitude of gene effect on outcome.
In order to detect genes which are conditionally related to survival, we considered the regularization patterns of the gene expression coefficients. The regularization pattern for a constrained coefficient is defined as the set of the estimated values of that coefficient varying the penalty parameter.
For comparison we used a $L1$ penalized Cox regression model, previously proposed for this kind of data [18]. The $L1$ penalized model introduces a constraint on the absolute value of the gene expression coefficients of the model (6):

$$\sum_{k=1}^{K} \mid \phi_k \mid < s. \quad 10]$$

The model can be fitted maximizing the penalized Partial Log-Likelihood:

$$\ell_\theta(\gamma_1, \gamma_2, \ldots, \gamma_m, \phi_1, \phi_2, \ldots, \phi_{30}) = \\ \ell(\gamma_1, \gamma_2, \ldots, \gamma_m, \phi_1, \phi_2, \ldots, \phi_K) - \frac{\theta}{2} \sum_{k=1}^{30} \mid \phi_k \mid \cdot \quad [11]$$

$L1$ penalized regression is usually considered as a method for variable selection, because by introducing the constraint (10), most of the estimated coefficients would be set exactly to zero.

The penalized Partial Log-Likelihood (8) was maximized using the algorithm implemented in the survival library of *R* software (*coxph* function with *ridge* option) [13]. The *L*1 penalized Cox regression model (11) was estimated using the algorithm proposed by Park and Hastie (2006) [18] and implemented in the *glmpath* library of *R* software. This algorithm uses the predictor-corrector method to determine the entire path of the coefficients's estimates as the penalty parameter varies. An application of this algorithm for the Cox regression model can be found in [19]. Other software like the *penalized* library of *R* software by J Goeman developed for penalized estimation in generalized linear models, support the Cox Proportional Hazard Model.
In the penalized regression models the gene expression values were re-scaled to have unit variance.

Simulation Study

We assessed the performance of the proposed method by a simulation study. The purpose was to show that the proposed approach selects the genes which are most related to survival under different correlation scenarios. We conducted simulations on both *L*1 and *L*2 penalized model to illustrate the different behavior of the two penalties.
Pseudo gene expression values were generated sampling from a multivariate normal distribution with the mean vector equal to 0 and exchangeable correlation structure:

$$\mathbf{Z} \sim N(0,\Sigma) \quad where\ \Sigma = \begin{pmatrix} 1 & \rho & \cdots \\ \rho & 1 & \vdots \\ \vdots & \cdots & \ddots \end{pmatrix} . \quad [12]$$

5000 covariates were generated under three different correlation scenarios, setting the correlation parameter equal to 0, 0.3 and 0.6. Since each gene has the same correlation with every other gene, the simulated scenario is much simpler than the real one; while in real data, only groups of genes are correlated with each other at different correlation levels.
The correlation patterns of real data are difficult to reproduce. In any case, this application on the simpler simulated scenario provides a check that the method is properly working and offers an example to show the trade off between error and variance (see results).
Pseudo survival time for the $i-th$ subject was simulated according to the following model:

$$t_i = exp(-\alpha - \sum_k \beta_k z_{ki} + \epsilon_i) \quad [13]$$

where $\varepsilon_i \sim N(0, 10)$. We set the coefficient for the $k-th$ gene equal to 0.3 for $1 \le k \le 5$, −0.3 for $6 \le k \le 10$ and 0 for $11 \le k \le 100$.
Censoring times were simulated sampling from a normal distribution $N(0, 10) + c$ where c is chosen to yield about a 20% censoring rate.
The sample size was set equal to 125. We analyzed each pseudo data set using both a *L*1 and a *L*2 approach after marginal pre-selection.

Results

ITT Colon Cancer Data

The penalized Cox regression model considered stage and location of the tumor, gender and age at time of surgery as relevant markers and the expression values of 60 candidate genes resulting from the preliminary selection (see Section 2.1). In Figure 3.1 the regularization patterns of the coefficients of the 60 top ranked genes from the *L*2 penalized Cox regression model are plotted. Increasing the penalty parameter θ, the estimated values of gene expression coefficients become close to zero. A small subgroup of genes (AGRP, UCK1, TBC1D7, FLJ22175, DEFB1, TXNL2, SURF2) is clearly separated from the rest of gene patterns [16]. Their coefficient estimates are negative indicating that a high gene expression value is associated to a reduced risk of death.
The *L*1 penalized Cox regression model gave similar results, (Figure 3.2). In particular four genes are detected by both methods.
The penalized Cox regression model considered stage of the tumor and age of patients as relevant markers and the expression values of 60 candidates genes resulting from the preliminary selection (see Section 2.1).
As for the ITT data, the stage of the tumor was the

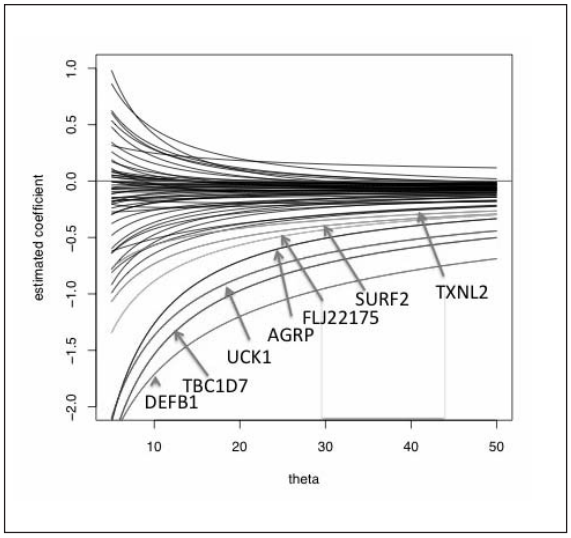

Figure 3.1. Regularization patterns of the coefficients of the 60 top ranked genes from the *L*2 penalized Cox Regression model for the ITT data. The coloured lines represent the coefficients for the following genes: DEFB1, TBC1D7, UCK1, AGRP, FLJ22175, SURF2, TXNL2.

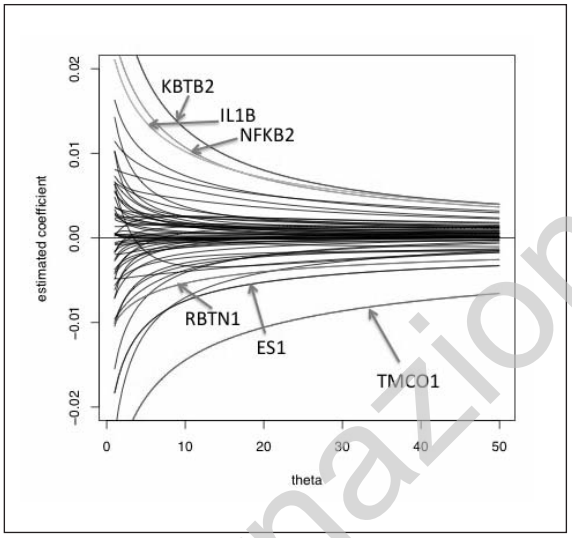

Figure 3.3. Regularization patterns of the 60 top ranked genes of the Batthacharjee's data from the *L*2 penalized Cox regression model. The coloured lines represent the coefficients of the following genes: RBTN1, ES1, TMCO1, IL1B, KBTB2, NFKB2.

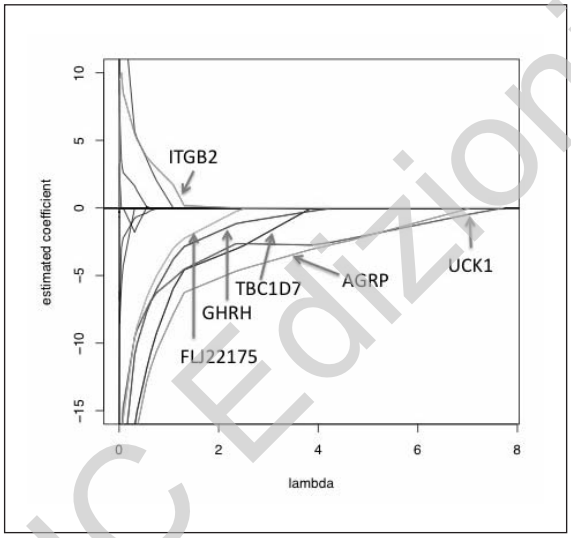

Figure 3.2. Regularization patterns of the coefficients of the 60 top ranked genes from the *L*1 penalized Cox regression model for the ITT data. The coloured lines represent the coefficients of the following genes: TBC1D7, UCK1, GHRH, FLJ22175, AGRP, ITGB2.

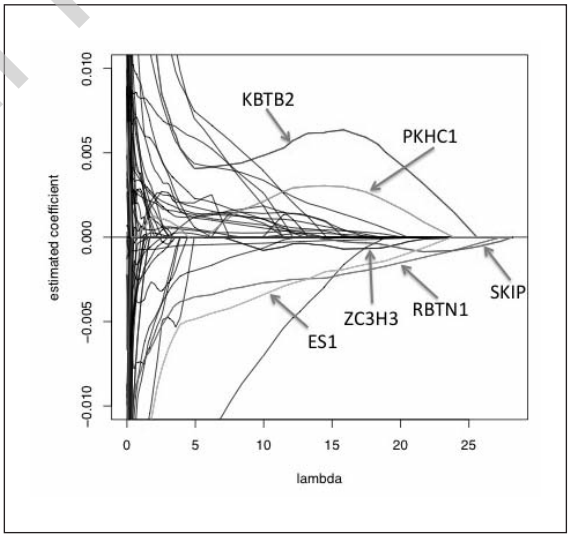

Figure 3.4. Regularization patterns of the coefficients of the 60 top ranked genes of the Batthacharjee's data from the L1 penalized Cox regression model. The coloured lines represent the coefficients of the following genes: RBTN1, ZC3H3, SKIP, ES1, KBTB2, PKHC1.

most relevant prognostic factor. Age also resulted in being significantly influential on survival (results not reported).

The results of the *L*2 penalized Cox regression model are reported in Figure 3.3. Six genes appear to be more related to survival: *RBTN1*, *ES1*, *TMCO1*, *IL1B*, *KBTB2*, *NFKB2*.

When we applied *L*1 penalization, we found similar results (Figure 3.4). Three genes were related to survival under both approaches.

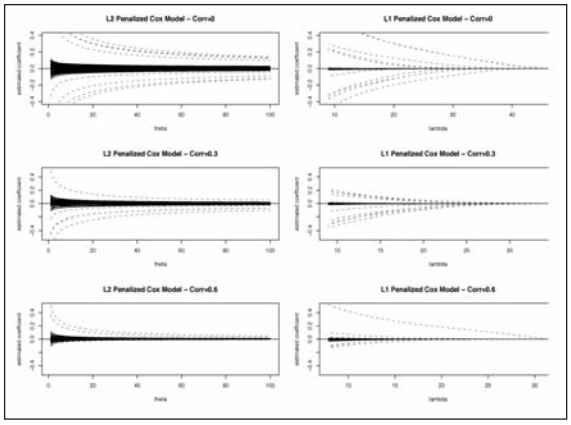

Figure 3.5. Averaged regularization patterns calculated over the 500 simulations. The dashed lines represent the genes with real values of the coefficients different from zero.

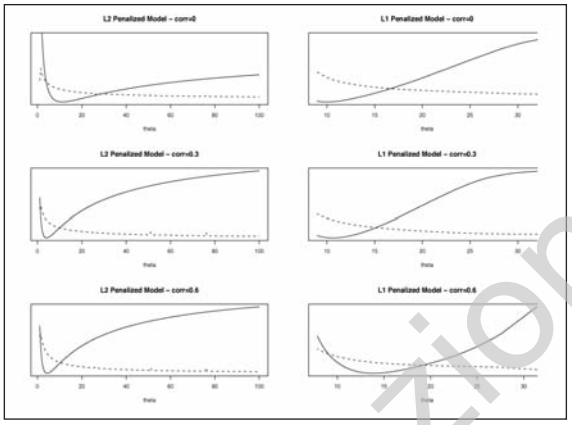

Figure 3.6. Bias (solid line) and variance (dashed line) of the regression coefficient estimator for one of the genes which are related to survival varying the smoothing parameter.

Simulation Studies

For each scenario, 500 runs were simulated. In Figure 3.5 we report for each gene the average regularization pattern calculated over the 500 simulations. The average regularization patterns of the genes which are related to survival are plotted as dashed lines. We can see that as the data correlation increases the separation between relevant and non relevant genes becomes less clear. However the dashed curves remain always the more external with respect to the others. For each value of the penalty parameter, we calculated the squared bias and variance of the estimator of each gene-expression coefficient:

$$(bias)^2 = (\bar{\beta} - \beta_0)^2 \qquad [15]$$

$$var = \frac{\sum(\beta_j - \bar{\beta})^2}{500} \qquad [16]$$

where β_0 is the true value of the gene expression coefficient, $\hat{\beta}_j$ is its estimated value under a specific value of the penalty parameter and $\bar{\beta}$ is the average of the estimated values over the 500 simulations.

In Figure 3.6, we show squared bias and variance of the regression coefficient estimator for one of the genes that are related to survival. We report results for $L1$ and $L2$ approach, under the three different correlation scenarios ($corr$ = 0, $corr$ = 0.3, $corr$ = 0.6). The trade off between error and variance is clearly evident. Even if the choice of a specific value for the smoothing parameter is not the aim of the present work, it should be noted that if the interest is in selecting θ, this can be done by considering the region where the bias is still of the same order of magnitude as the variance. In this way we control the variance-bias trade off.

Discussion

Biological Significance of ITT Data Analysis Results

According to the present analysis there are four genes discovered by both methods as associated with survival. Two out of four might have a direct role in better response to chemotherapy. FLJ22175 induces chromosomal instability as well as hypersensitivity to DNA cross-linking agents including a number of chemotheutic agents. UCK1 is an homologous UCK2, reported to play a crucial role in activating anti-tumor pro-drugs in human cancer cells [22].

The other two genes TBC1D7 and AGRP do not have a clear association to cancer, but are relevant for fundamental processes important for the susceptibility to cancer. TBC1D7 is involved in G-protein signal transduction. Another gene of this class GHRH, reported expressed in pituitary tumors and Adenomas, has been discovered to be associated with good prognosis with the $L1$ penalized model, but not with the $L2$ model.

AGRP may play a role in the central control of fee-

ding. Its overexpression could be associated with fat infiltration and inflammatory response. Another gene associated in survival by $L2$ analysis is DEFB1, a gene controlling the innate immune response to bacterial infection. Loss or underexpression of DEF1B has been reported in renal and prostatic carcinomas [8]. DEFB1 overexpression has been reported to induce apoptosis and this gene is proposed to be a tumor suppressor [23]. ITGB2, gene involved in leukocyte migration during inflammatory response, was discovered by $L1$ analysis to be associated with poor survival [6].

Biological Significance in Bhattacharjee's Data Analysis Results

Among the genes that we discovered from the $L2$ analysis one can notice IL1B (produced by activated macrophages, inducing IL-2 release, B-cell maturation) which could be a sign of macrophages or B cells infiltration of the tumor.
A second important candidate, NFKB2, is also involved in the immune response. These findings could indicate that activation of the inflammatory response is a marker of gravity and enhanced tumor progression, as supported also by classic tumor classification and the analysis performed on colon cancer (see above). Among the other genes discovered as important, KBTB2 and TMCO1 are so far poorly characterized, but reported to be expressed in the lung.
Two genes, RBTN1 (from both the analyses) and ZC3H3 (from $L1$ only), not previously associated with lung cancer could play an important role and are reported to be transcriptional activators.
Considering the larger list of 60 genes included in the penalized models, we should notice the presence of BMP7 and of ASCL1, part of Cluster C2 also associated with survival in the paper from Bhattacharjee [4]. It is worth noticing that a gene of the BMP family, BMP6, was previously reported [15] as being associated with survival among the squamoid tumors.

Methodologic Issues

In this study we analyzed the prognostic value of gene expression in predicting survival of cancer patients using a penalized regression approach. We applied this method on two different data sets. The advantage of this approach is to estimate the net contribution of differential gene expression, given other relevant prognostic variables and the conditional relationship among genes. We applied a two step procedure. After having pre-selected the genes which are mostly related to survival by an univariate analysis, we applied $L2$ penalized regression.
Our method intends to evaluate whether or not a gene has an effect on survival. This can be seen as a test for trend.
We are not interested here in modeling the dose-response curve. In literature we find several example of fully non parametric methods like kernel machine or principal value decomposition [10]. The preliminary selection, performed as a first step, is intended to reduce the complexity of the penalized model. In literature there are examples of models where pre-selection cannot be avoided, as we can see in [3] where they rank genes using Cox Proportional Hazards Model. This pre-selection is obviously critical because it can bring to biased results. First, it precludes any chance of detecting genes that are conditionally but not marginally related to survival.
Second, we adjusted the effect of each pre-selected gene for the other pre-selected genes only and not for the entire gene expression matrix. This can increase the number of false negatives and false positives in a way that can not be simply quantified.
In principle we could evaluate the impact of pre-selection on the final results by permutation. However applying permutation methods is not trivial in this context because of the presence of covariates. The problem is in reproducing by permutations the distribution under the null hypothesis of partial random association. While solutions are proposed for linear models [2], use of permutation tests in censored survival data analysis is not largely discussed in literature and permutation tests in a multivariate Cox model is an open issue.
A partial answer to pre-selection consequences comes from the $L1$. In fact, we should note that $L1$ penalty allows the inclusion of a higher number of covariates in the model because the stronger con-

straint allows one to obtain stable results even if the entire set of genes is included in the predictor. To evaluate if pre-selection affects results, we applied the $L1$ penalized model to a larger group of genes (120 genes) and to the entire gene expressions matrix (2587 genes).
The results were consistent with the ones obtained after pre-selection (results not reported). This fact indicates that in our example pre-selection does not largely affect results and that the choice of K = 60 is not sensitive.
In literature, several methods have been proposed for automatic selection of the regularization parameter in $L1$ and $L2$ penalized regression. They are usually based on Cross Validation, Generalized Cross Validation, bootstrap or algorithms like the elastic net [24] [29]. In our application no choice of the penalty parameter was done, but relevant genes were explored looking at the regularization patterns [16][28]. For several reason we prefer to look at the entire regularization patterns instead of selecting the smoothing parameter. First, penalized regression can be used when the number of covariates is larger than the sample size. This is an ill posed problem which requires some additional assumptions in order for a stable solution to be obtained. From a Bayesian point of view, these additional assumptions express our prior belief about the problem. In our particular situation, each specific value of the regularization parameter corresponds to a specific prior belief about the gene expression coefficients. Looking at the entire regularization pattern, we are not super-imposing any specific prior on genes coefficients, expressing our ignorance about the predictive value of gene expression on survival. The natural extension of our approach is a fully Bayesian penalized model which combines data information and prior belief on the regularization parameter in a posterior joint distribution, which has a model averaging interpretation. Adapting an informal Bayesian method, the researcher can choose the preferred range of the penalty parameter. For example, relying on the fact that small values of the coefficients are more realistic, one can give more importance to the right region of the regularization curves where the penalty parameter has larger values.
Secondly, regularizations patterns are very informative. In principle, they can be used to understand the complex interplay among covariates [16]. Once the regularization patterns are reported, selecting the smoothing parameter via a smoothing parameter selector (such as GCV) would not give any additional information. Moreover, it is known that standard practices based on cross-validation tend to select overparametrized models, while criteria like BIC tend to prefer less complex models [21]. This indicates that each specific selector has an underlying prior assumption, but it is not explicitly specified.
The regularization pattern gives a two-dimensional portrait of the effect of correlation among covariates [16]. For example it can happen that a covariate has the negative coefficient with the largest absolute value when $\theta = 0$, but the increase of the smoothing parameter quickly drives it toward zero becoming even positive. This could be explained by the fact that this covariate has a strong correlation with another one and they are stable as a sum. This behaviour is not atypical, especially when the covariates are correlated to various degrees with other factors. So looking at the entire regularization patterns gives us an idea of the complex correlation structure among genes.
For comparison purposes, we specified and fitted also a $L1$ penalized Cox regression model because it is a popular approach in expression data analysis [5]. Under the $L1$ model, increasing the regularization parameter, for the most part, the estimated coefficients become exactly zero. Therefore the reader should notice that fixing the smoothing parameter is equivalent to selecting a ”best” subset of predictors.
This application shows that even if in both applications a common subset of genes can be detected, $L2$ and $L1$ penalized Cox regression models give similar but not exactly equal results. Discrepancies are explained by the fact that the two methods rely on different assumptions and their performance depends on data characteristics. Statistical properties of $L1$ and $L2$ penalized regression models and differences between the two penalty approaches already have been explored [25] [24] [14]. Here we can recall known characteristics of the two approaches which can be useful for interpreting the results of the penalized regression and, as a consequence, can motivate a method choice.

From a Bayesian point of view, it can be shown that *L*2 penalized regression assumes a Gaussian prior distribution for the coefficients, while *L*1 penalized regression assumes a double exponential prior distribution [11]. The double exponential distribution produces more mass near zero and in the tails. This is the reason why under the *L*1 approach coefficients's estimates are either larger than zero or zero than under *L*2 approach [12]. Ridge regression tends to retain small parameters. This implies that if the true model includes many small but non-zero regression parameters, *L*1 will perform poorly while *L*2 will perform well. If the true model includes many zero parameters, *L*1 will outperform *L*2 [24].
Regarding data characteristics, we have to consider multi-collinearity among covariates. In fact, the amount of shrinkage of the coefficients varies under *L*1 or *L*2 penalization, depending on multi-collinearity [24]. Under the *L*2 penalty, groups of correlated variables can be selected together [29]. On the contrary, given a set of highly correlated variables associated with outcome, procedures that employ a penalty function that is not strictly convex, like the *L*1 penalty, often will identify only one of the variables and ignore the others [24].

Conclusion

The penalized Cox regression model is a useful tool in genomics data analysis when we want to estimate the predictive value of gene expression on survival while taking into account for other prognostic variables. The *L*2 penalized regression approach allows to simultaneously modeling of the contribution of several candidate genes.
Some considerations indicate the *L*2 penalized regression as a more appropriate approach in this context than *L*2.
The tendency of the *L*1 penalty to identify only one gene within a group of correlated variables is a limitation in the analysis of gene expression data where the identification of a whole set of correlated genes may lead to an improved understanding of the biological pathway. This point should be taken into account in interpreting the regression results.
Moreover, in presence of strong correlation, *L*1 penalization can bias effect estimates. Indeed, when one gene coefficient is set exactly to zero, the coefficients of the correlated genes are no longer adjusted by his presence and can suddenly grow or decrease. We can say that when one gene is excluded from the model, we do not adjust for it. As the regularization parameter increases, the conditions under which we estimate the model change. On the contrary, in the *L*2 penalized regression approach, we always adjust for all the rest of the genomic information, since every genes is always present: even if shrunken toward zero the coefficient never leaves the model.
Regarding the choice of the regularization parameter, we propose to avoid selection and consider the entire regularization patterns varying the penalty parameter. A future step would be to explicitly insert a prior distribution on the penalty parameter. The natural extension of our approach is a fully Bayesian penalized model which combines data information and prior belief on the regularization parameter.

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage WR J O, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000,403:503.
2. Anderson MJ, Legendre P: An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Statist. Comput. Simul. 1999, 62:271–303.
3. Annest A, Bumgarner RE, Raftery AE, Yeung KY: Iterative Bayesian model averaging: a method for the application of survival analysis to high-dimensional microarray data. BMC Bioinformatics 2009, 10:72.
4. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M,Weber G, Mark EJ, Lander ES,WongW, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA 2001, 98(24):13790–13795; (www.broad.mit.edu/mpr/lung).
5. Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, Frigessi A, Lingjaerde OC: Predicting survival-

from microarray data - a comparative study. Bioinformatics 2007, 23(16):2080–2087.
6. Cavalieri D, Dolara P, Mini E, Luceri C, Castagnini C, Toti S, Maciag K, DeFilippo C, Nobili S, Moranti M, Napoli C, Tonini G, Baccini M, Biggeri A, Tonelli F, Valanzano R, Orlando C, Gelmini S, Cianchi F, Tesserini L, Luzzatto L:Analysis of gene expression profiles reveals novel correlations with the clinical course of colorectal cancer. Oncology research 2007, 16:535-548.
7. Cox DR: Regression models and life tables. J. Royal Statis. Soc. 1972, 34:187–220.
8. Donald C, Sun C, Lim S, Macoska J, Cohen C, Amin M, Young A, Ganz T, Marshall F, Petros J: Cancer-specific loss of beta-defensin 1 in renal and prostatic carcinomas. Lab Invest. 2003, 83(4):501–5.
9. Draper N, Smith H: Applied regression analysis. Wiley and Sons Ltd 1981.
10. Dupuy A, Simon M: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007, 99(2):147–157.
11. Fu W: penalized regression: the Bridge versus the Lasso. Journal of Computational and Graphical Statistics 1998, 7(3):397–416.
12. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC: Testing association of a pathway with survival using gene expression data. Bioinformatics 2005, 21(9).
13. Gray: Flexible methods of analysing survival data using splines with applications to breast cancer prognosis. J of the American Stat. Ass. 1992, 87(420).
14. Gui J, Li H: penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Annals of statistics 2004, 32(2):407–499.
15. Hayes D, Monti S, Parmigiani G, Gilks C, Naoki K, Bhattacharjee A, Socinski M, Perou C, Meyerson M: Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohort. Journal of Clinical Oncology 2006, 24(31):5079–89.
16. Hoerl AE, Kennard RW: Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970, 12:55–67.
17. Lokhorst J: The lasso and generalized linear models. Technical report 1999.
18. Park MY, Hastie T: L1 regularization algorithm for generalized linear models. Technical report 2006.
19. Park MY, Hastie T, Tibshirani R: Averaged gene expressions for regression. Biostatistics 2007, 8(2):212–227.
20. Paul D, Bair E, Hastie T, Tibshirani R: Pre-conditioning for feature selection and regression in highdimensional problems. Tech report April 2006;(http://www-stat.stanford.edu/ tibs/ftp/precond.pdf).
21. Ruppert D, Wand M, Carroll R: Semiparametric regression. Cambridge University Press 1999.
22. Shimamoto Y, Koizumi K, Okabe H, Kazuno H, Murakami Y, Nakagawa F, Matsuda A, Sasaki T, Fukushima M: Sensitivity of human cancer cells to the new anticancer ribo-nucleoside TAS-106 is correlated with expression of uridine-cytidine kinase 2. Jpn J Cancer Res. 2002, 93(7):825–33.
23. Sun C, Arnold R, Fernandez-Golarz C, Parrish A, Almekinder T, He J, Ho S, Svoboda P, Pohl J, Marshall F, Petros J: Human beta-defensin-1, a potential chromosome 8p tumor suppressor: control of transcription and induction of apoptosis in renal cell carcinoma. Cancer Research 2006, 66(17):8542–9.
24. Tibshirani R: regression shrinkage and selection via the LASSO. J Royal Statist. Soc. Series B 1996, 58:267–288.
25. Tibshirani R: The LASSO method for variable selection in the Cox model. Statistics in Medicine 1997, 16(4):385–395.
26. Tusher V, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. PNAS 2001, 98:5116–5121.
27. van Wieringen WN, Kun D, Hampel R, Boulesteix AL: Survival prediction using gene expression data: a review and comparison. Computational statistics and data analysis 2008, 23(16):2080–2087.
28. Zhang R, McDonald G: Characterization of ridge trace behavior. Communications in Statistics-Theory and Methods 2005, 34(7):1487–1501.
29. Zou H, Hastie T: Regularization and variable selection via the elastic net. Journal of Royal Statistical Society 2005, 67(2):301–320.

Appendices

Details on Score Test Calculation

In our context we need to consider many extensions to a model adding one out of many new explanatory variables (genes). Doing this we may wish to avoid maximizing the Likelihood under each extended model. This is what Rao's score tests do, by approximating the log-likelihood from its shape at β_0. Indicating with $\eta = (\gamma_i, \beta)$ the vector of the unknown parameters, the partial likelihood can be written as:

$$L(\eta) = \prod_{i=1}^{M} \left(\frac{exp(\eta^{\mathbf{T}} \mathbf{Z}_i)}{\sum_{j \in R_i} exp(\eta^{\mathbf{T}} \mathbf{Z}_j)} \right) \quad [16]$$

where $\mathbf{Z}$ is the vector of the expressions of the covariates considered in the model for the ith subject, M gives the number of the subjects who experience the event and R_i defines the risk set when the ith event happens.
Weighting each subject according to his relative risk, we can now define the mean vectors $\overline{Z}_i(\eta)$, $i = 1, \ldots, M$

$$\hat{Z}(\eta) = \frac{\sum_{j \in R_i} \mathbf{Z}_j \exp(\eta^{\mathbf{T}} \mathbf{Z}_j)}{\sum_{j \in R_i} \exp(\eta^{\mathbf{T}} \mathbf{Z}_j)} \quad [17]$$

and the variance vectors $Var_i^Z(\eta)$

$$Var_i^Z(\eta) = \frac{\sum_{j \in R_i} \mathbf{Z}_j^2 exp(\eta^{\mathbf{T}} \mathbf{Z_j})}{\sum_{j \in R_i} exp(\eta^{\mathbf{T}} \mathbf{Z_j})} - (\overline{Z}(\eta))^2 \quad [18]$$

The partial log-likelihood is defined as

$$\ln L(\eta) = \sum_{i=1}^{M} \left(\eta^{\mathbf{T}} \mathbf{Z_i} - \ln \sum_{j \in R_j} exp(\eta^{\mathbf{T}} \mathbf{Z_j}) \right) \quad [19]$$

so the score vector is

$$\mathbf{U}(\eta) = \sum_{i=1}^{M} (Z_i - \hat{Z}_i(\eta)) \quad [20]$$

and the observed information:

$$I(\eta) = -\sum_{i=1}^{M} Z Var_i^Z(\eta) \quad [21]$$

The score vector under the null hypotesis $\beta = 0$ is given by

$$\mathbf{U}(\eta)^T|_{(0,\hat{\gamma}_1,\ldots,\hat{\gamma}_p)} = [U_g, 0..., 0] \quad [22]$$

where $\hat{\gamma}_1, ..., \hat{\gamma}_p$ are the partial maximum likelihood estimates of the parameters under the model without the gene effect ($\beta = 0$), and U_g is the first element of the $\mathbf{U}(\eta)$ vector, calculated in $\beta = 0$. Writing $J = [J_{ij}] = [I(\eta)]|_{(0,\hat{\gamma}_1,\ldots,\hat{\gamma}_p)}$, the test statistic can be written as

$$T = [U_g, 0..., 0] J^{-1} [U_g, 0..., 0] = J_{11}^* U_g^2 \quad [23]$$

where J_{11} is the element (1, 1) of the inverse matrix J^{-1}.

Acknowledgement

The present research was made possible by support from: PRIN-2005134079 "Statistical tools in system biology", European Union VI Framework, Network of Excellence NuGO and by ITT, Florence, Italy.