Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 21

Univariate Shrinkage in the Cox Model for High Dimensional Data

Robert J. Tibshirani*

*Stanford University, tibs@stat.stanford.edu

Univariate Shrinkage in the Cox Model for High Dimensional Data*

Robert J. Tibshirani

Abstract

We propose a method for prediction in Cox's proportional model, when the number of features (regressors), p, exceeds the number of observations, n. The method assumes that the features are independent in each risk set, so that the partial likelihood factors into a product. As such, it is analogous to univariate thresholding in linear regression and nearest shrunken centroids in classification. We call the procedure Cox univariate shrinkage and demonstrate its usefulness on real and simulated data. The method has the attractive property of being essentially univariate in its operation: the features are entered into the model based on the size of their Cox score statistics. We illustrate the new method on real and simulated data, and compare it to other proposed methods for survival prediction with a large number of predictors.

KEYWORDS: proportional hazards model, survival data, lasso, high-dimensional

---

*I thank Daniela Witten for helpful comments, and acknowledge support from National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

1 Introduction

High-dimensional regression problems are challenging, and a current focus of statistical research. One simplifying assumption that can be useful is that of independence of features. While this assumption will rarely be even approximately true, in practice, it leads to simple estimates with predictive performance sometimes as good or better than more ambitious multivariate methods.

In linear regression with squared error loss and a lasso penalty, independence of the features leads to the *univariate soft-thresholding* estimates, that often behave well.

In classification problems, the idea of class-wise independence forms the basis for diagonal discriminant analysis. In that setting, if we assume that the features are independent within each class, and incorporate a lasso penalty, the resulting estimator is the *nearest shrunken centroid* procedure proposed by Tibshirani et al. (2001): we review this fact in the next section. The nearest shrunken centroid classifier is very simple but useful in practice, and has many attractive properties. It has a single tuning parameter that automatically selects the features to retain in the model, and it provides a single ranking of all features, that is independent of the value of the tuning parameter.

We review the details of the regression and classification problems in the next section. Then we extend the idea of class-wise feature independence to the Cox model for survival data: this is the main contribution of the paper. We study this new procedure on both real and simulated datasets, and compare it to competing methods for this problem.

2 Regression and classification using feature independence

2.1 Regression

Consider the usual regression situation: we have data $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots n$ where $\mathbf{x}_i = (x_{i1}, \ldots x_{ip})^T$ and y_i are the regressors and response for the ith observation. Assume the x_{ij} are standardized so that $\sum_i x_{ij}/n = 0, \sum_i x_{ij}^2/n = 1$. The lasso finds $\beta = (\beta_1, \ldots \beta_p)^T$ to minimize

$$\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1}$$

Now if we assume that the features are independent (uncorrelated), it is

easy to show that

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^o)(|\hat{\beta}_j^o| - \lambda)^+ \tag{2}$$

where $\hat{\beta}_j^o$ are the ordinary (univariate) least squares estimates. These are *univariate soft-thresholded estimates*, and perform particularly well when $p \gg n$. Zou & Hastie (2005) study these estimates as a special case of their elastic net procedure. See Fan & Lv (2008) for some theoretical support of univariate soft-thresholding.

2.2 Classification

Here we review the nearest shrunken centroid method for high-dimensional classification and its connection to the lasso. Consider the *diagonal-covariance* linear discriminant rule for classification. The *discriminant score* for class k is

$$\delta_k(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}_{kj})^2}{s_j^2} + 2\log \pi_k. \tag{3}$$

Here $x^* = (x_1^*, x_2^*, \ldots, x_p^*)^T$ is a vector of expression values for a test observation, s_j is the pooled within-class standard deviation of the jth gene, and $\bar{x}_{kj} = \sum_{i \in C_k} x_{ij}/n_k$ is the mean of the n_k values for gene j in class k, with C_k being the index set for class k. We call $\tilde{x}_k = (\bar{x}_{k1}, \bar{x}_{k2}, \ldots \bar{x}_{kp})^T$ the *centroid* of class k. The first part of (3) is simply the (negative) standardized squared distance of x^* to the kth centroid. The second part is a correction based on the class *prior probability* π_k, where $\sum_{k=1}^{K} \pi_k = 1$. The classification rule is then

$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \max_k \delta_k(x^*). \tag{4}$$

We see that the diagonal LDA classifier is equivalent to a nearest centroid classifier after appropriate standardization.

The nearest shrunken centroid classifier regularizes this further, and is defined as follows. Let

$$d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0)}, \tag{5}$$

where $\bar{x}_j$ is the overall mean for gene j, $m_k^2 = 1/n_k - 1/n$ and s_0 is a small positive constant, typically chosen to be the median of the s_j values. This constant guards against large d_{kj} values that arise from expression values near zero.

Finally, we can derive the nearest shrunken centroid classifier as the solution to a lasso-regularized problem that assumes independence of the features. Consider a (naive Bayes) Gaussian model for classification in which

DOI: 10.2202/1544-6115.1438

the features $j = 1, 2, \ldots, p$ are assumed to be independent within each class $k = 1, 2, \ldots, K$. With observations $i = 1, 2, \ldots, n$ and C_k equal to the set of indices of the n_k observations in class k, we observe $x_{ij} \sim N(\mu_j + \mu_{jk}, \sigma_j^2)$ for $i \in C_k$ with $\sum_{k=1}^K \mu_{jk} = 0$. We set $\hat{\sigma}_j^2 = s_j^2$, the pooled within-class variance for feature j, and consider the lasso-style minimization problem

$$\min_{\{\mu_j, \mu_{jk}\}} \left\{ \frac{1}{2} \sum_{j=1}^{p} \sum_{k=1}^{K} \sum_{i \in C_k} \frac{(x_{ij} - \mu_j - \mu_{jk})^2}{s_j^2} + \lambda \sqrt{n_k} \sum_{j=1}^{p} \sum_{k=1}^{K} |\frac{\mu_{jk}|}{s_j} \right\}. \tag{6}$$

By direct calculation, it is easy to show that the solution is equivalent to the nearest shrunken centroid estimator (5), with s_0 set to zero, and M_k equal to $1/n_k$ instead of $1/n_k - 1/n$ as before. Some interesting theoretical results for diagonal linear discriminant analysis are given in Bickel & Levina (2004). Efron (2008) proposes an Empirical Bayes variation on nearest shrunken centroids.

3 Survival data and Cox's proportional hazards model

For the rest of this paper we consider the right-censored survival data setting. The data available are of the form $(y_1, \mathbf{x}_1, \delta_1,), \ldots, (y_n, \mathbf{x}_n, \delta_n)$, the survival time y_i being complete if $\delta_i = 1$ and right censored if $\delta_i = 0$, with $\mathbf{x}_i$ denoting the usual vector of predictors $(x_1, x_2, \ldots x_p)$ for the ith individual. Denote the distinct failure times by $t_1 < \cdots < t_K$, there being d_k failures at time t_k.

The proportional-hazards model for survival data, also known as the Cox model, assumes that

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\Big(\sum_j x_j \beta_j\Big) \tag{7}$$

where $\lambda(t|\mathbf{x})$ is the hazard at time t given predictor values $\mathbf{x} = (x_1, \ldots, x_p)$, and $\lambda_0(t)$ is an arbitrary baseline hazard function.

One usually estimates the parameter $\beta = (\beta_1\ \beta_2\ \ , . . \beta_p)^T$ in the proportional-hazards model (7) without specification of $\lambda_0(t)$ through maximization of the partial likelihood

$$\mathrm{PL}(\beta) = \prod_{k \in D} \frac{\exp(\mathbf{x}_k \beta)}{\{\sum_{m \in R_k} \exp(\mathbf{x}_m \beta)\}}.$$

Here D is the set of indices of the failure times, R_k is the set of indices of the individuals at risk at time $t_k - 0$, and we have assumed there are no ties in

the survival times. In the case of ties, we use the "Breslow" approximation to the partial likelihood. We also assume that the censoring is noninformative, so that the construction of the partial likelihood is justified.

If $p \gg n$ then maximizer of the partial likelihood is not unique, and some regularization must be used. One possibility is to include a "lasso" penalty and jointly optimize over the parameters, as suggested by Tibshirani (1997).

Here we extend the idea of feature independence to the Cox model for survival data. Now the partial likelihood term $\exp(\mathbf{x}_k\beta)/\sum_{m\in R_k}\exp(\mathbf{x}_m\beta)$ equals $\Pr(k|\mathbf{x}_i, i \in R_k)$ the probability that individual k is the one who fails at time t_k, given the risk set and their feature vectors. Now we assume that both conditionally on each risk set, and marginally, the features are independent of one another. Then using Bayes theorem we obtain $\Pr(k|\mathbf{x}_i, i \in R_k) \sim \prod_j \Pr(k|x_{ij}, i \in R_k)$. As a result, up to a proportionality constant, the partial likelihood becomes a simple product

$$\mathrm{PL}(\beta) = c \cdot \prod_{j=1}^{p}\prod_{k\in D}\frac{\exp(x_{kj}\beta_j)}{\{\sum_{m\in R_k}\exp(x_{mj}\beta_j)\}}$$

The log partial likelihood is

$$\begin{aligned}\ell(\beta) &= \sum_{j=1}^{p}\sum_{k=1}^{K}\big(x_{kj}\beta_j - \log\sum_{m\in R_k}\exp(x_{mj}\beta_j)\big) &(8)\\ &\equiv \sum_{j=1}^{p} g_j(\beta_j). &(9)\end{aligned}$$

We propose the *Cox univariate shrinkage* (CUS) estimator as the maximizer of the penalized partial log-likelihood

$$J(\beta) = \sum_{j=1}^{p} g_j(\beta_j) - \lambda\sum|\beta_j|. \qquad (10)$$

Here $\lambda \geq 0$ is a tuning parameter, and we solve this problem for a range of λ values. Note that is just a set of one-dimensional maximizations, since we can maximize each function $g_j(\beta_j) - \lambda|\beta_j|$ separately. The minimizers of (10) are not simply soft-thresholded versions of the unpenalized estimates $\hat{\beta}_j$, as they are in the least squares regression case. We discuss and illustrate this fact in the next section. From this, we obtain solutions paths $(\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda), \ldots \hat{\beta}_p(\lambda))$, and we can use these to predict the survival in a separate test sample.

Note that the estimates $\hat{\beta}$ will tend to be biased towards zero, especially for larger values of λ. In the test set and cross-validation computations in this

DOI: 10.2202/1544-6115.1438

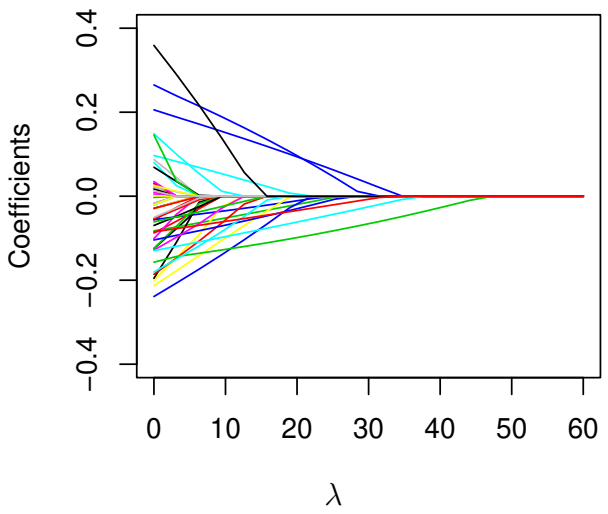

Figure 1: *Kidney cancer example: solution paths for univariate shrinkage, for 500 randomly chosen genes.*

paper, we fit the predictor $z = \mathbf{x}\hat{\beta}$ in the new (test or validation) data, and so obtain a scale factor to debias the estimate. If we are interested in the values of $\hat{\beta}$ themselves, a better approach would be fit the single predictor $\gamma \mathbf{x}_i \hat{\beta}$ in a Cox model on the training data, and hence obtain the adjusted estimates $\hat{\gamma}\hat{\beta}$.

The lasso penalty above might not make sense if the predictors are in different units. Therefore we first standardize each predictor x_j by s_j, the square root of the observed (negative) Fisher information at $\beta_j = 0$.

3.1 Example

Zhao et al. (2005) collected gene expression data on $14,814$ genes from 177 kidney patients. Survival times (possibly censored) were also measured for each patient. The data were split into 88 samples to form the training set and the remaining 89 formed the test set. Figure 1 shows the solution paths for Cox univariate shrinkage for 500 randomly chosen genes.

Notice that the solutions paths have a particularly nice form: they are monotonically decreasing in absolute value and once they hit zero, they stay at zero. It is easy to prove that this holds in general. The proof is given in

the Appendix.

Note also that the profiles decrease at different rates. This is somewhat surprising. In the least squares regression case under orthogonality of the predictors, the lasso estimates are simply the soft-thresholded versions of the least squares estimates $\hat{\beta}_j$:, that is, $\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - t)_+$. Thus every profile decreases at the same rate. Here, each profile decreases at rate $1/|g'(\hat{\beta}_j)|$, which is the reciprocal of the observed Fisher information.

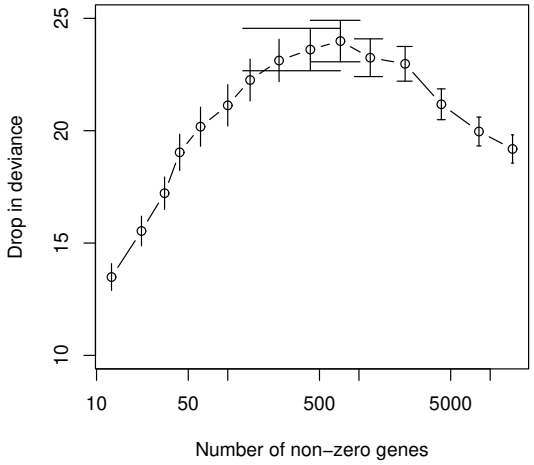

Figure 2: *Kidney cancer example: cross-validation curve. Drop in deviance on validation set, with plus and minus one standard error bars.*

Let U_j, V_j be the gradient of the (unpenalized) log-partial likelihood and the (negative) observed Fisher information, evaluated at $\beta_j = 0$ Then we can also show that

$$\hat{\beta}_\lambda \neq 0 \iff |U_j|/\sqrt{V_j} > \lambda. \tag{11}$$

The proof is given in the Appendix. Therefore we have the following useful fact:

For any value of λ, the set of predictors that appear in the model estimated by Cox univariate shrinkage are just those whose score statistic exceeds λ in absolute value.

That is, the Cox univariate shrinkage method used a fixed ranking of all features based on the Cox score statistic. This is very convenient for interpretation, as the Cox score is often used for determining the univariate significance

DOI: 10.2202/1544-6115.1438

of features (e.g. in the SAM procedure of Tusher et al. (2001)). However the non-zero coefficients given to these predictors are not simply the score statistics or soft thresholded versions of them.

Figure 3 shows the drop in test sample deviance for CUS (Cox univariate shrinkage), lasso, SPC (Supervised principal components) and UST (univariate soft thresholding), plotted against the number of genes in the model. The CUS and SPC methods work best.

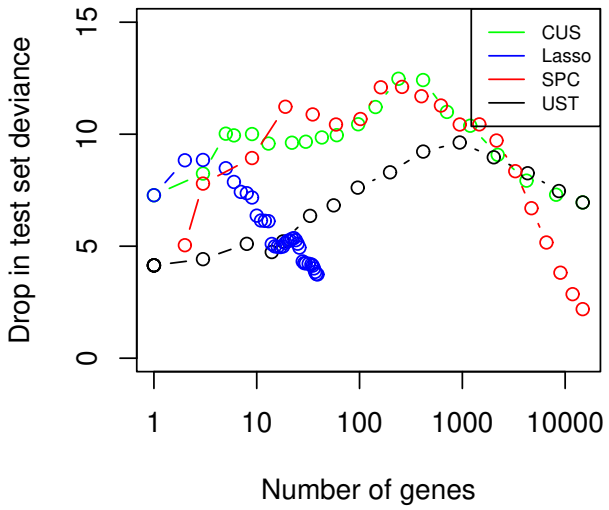

Figure 3: *Kidney cancer example: drop in test sample deviance for CUS (Cox univariate shrinkage), lasso, SPC (Supervised principal components) and UST (univariate soft thresholding).*

Figure 4 shows the univariate Cox coefficients plotted against the Cox score statistics, for the kidney cancer data. Although the two measures are correlated, this correlation is far from perfect, and again this underlines the difference between our proposed univariate shrinkage method, which enters features based on the size of the Cox score, and simple soft-thresholding of the univariate coefficients. The latter ranks features based on the size of their *unregularized* coefficient, while the former looks at the standardized effect size at the null end of the scale.

To estimate the tuning parameter λ we can use cross-validation. However it is not clear how to apply cross-validation in the setting of partial likeli-

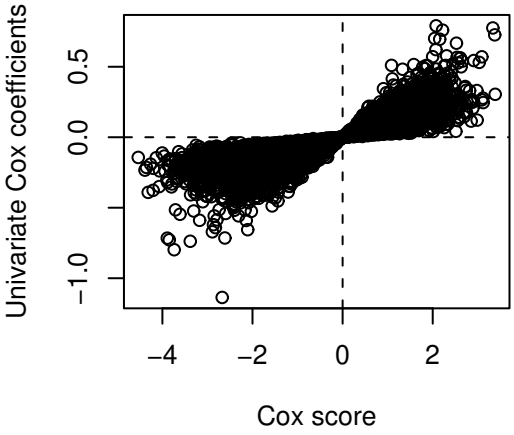

Figure 4: *Kidney cancer data: the univariate Cox coefficients plotted against the Cox score statistics.*

hood, since this function is not a simple sum of independent terms. Verweij & van Houwelingen (1993) propose an interesting method for leave-one-out cross-validation of the partial likelihood: for each left-out observation they compute the probability of its sample path over the risk sets. For the fitting methods presented in this paper, this leave-one-out cross-validation was too slow, requiring too many model fits. When we instead tried 5 or 10- fold partial likelihood cross-validation, the resulting curve was quite unstable. So we settled on a simpler approach, 5-fold cross-validation where we fit a Cox model to the left out one-fifth and record the drop in deviance. The results for the kidney cancer data is shown in Figure 2. The curve is a reasonable estimate of the test deviance curve of Figure 3.

In some microarray datasets, there are duplicate genes, or multiple clones representing the same gene. In this case, the data for the different features will be highly correlated, and it makes sense to average the data for each gene, before applying UC. One can also average genes whose pairwise correlation is very high, whether or not they're supposed to represent the same gene. For example, when we averaged genes with correlation greater than 0.9 in the kidney cancer dataset, it reduced the number of genes by a few hundred and caused very little change in the test deviance curve.

DOI: 10.2202/1544-6115.1438

3.2 Selection of a more parsimonius model

The CUS procedure enters predictors into the model based on this univariate Cox scores. Thus if two predictors are strongly predictive and correlated with each other, both will appear in the model. In some situations it is desirable pare down the predictors to a smaller, mode independent set. The lasso does this selection automatically, but sometimes doesn't predict as well as other methods (as in Figure 3.) An alternative approach is *pre-conditioning* (Paul et al. 2008), in which we apply the standard (regression) lasso to the fitted values from a model. In Paul et al. (2008) the initial fitting method used was supervised principal components. Here we start with the CUS fit using 500 genes, which approximately the best model from Figure 2. We then apply the (regression version) of the lasso to the fitted values, and we report the drop in test deviance in Figure 5, along with that for CUS itself (green curve copied from Figure 3.) We see that pre-conditioning gives about the same drop in test deviance as the 500 gene CUS model, but using fewer than 100 genes. And it performs better here than the Cox model lasso (blue curve in Figure 3, a finding supported theoretically in Paul et al. (2008). Figure 6 shows the univariate Cox scores for the predictors entered by each method. We see that the pre-conditioning approach enters only predictors that are individually strong, while the usual lasso enter many predictors that are individually weak.

4 Simulation study

We generated Gaussian samples of size $n = 50$ with $p = 1000$ predictors, with c population correlation $\rho(x_i, x_j)$ between each pair of predictors. The outcome y was generated as an exponential random variable with mean $\exp(\sum_1^{1000} x_j\beta_j)$, with a randomly chosen s out of the p coefficients equal to value ± 4 and the rest equal to zero. We considered two scenarios— $s = 20$ and $s = 200$—, and tried both $\rho(x_i, x_j) = 0$ and $\rho(x_i, x_j) = 0.5^{|i-j|}$. Three fitting methods for the Cox model were compared: lasso, supervised principal components, and CUS, over a range of the tuning parameter for each method. A test set of size 400 was generated and the test set deviance of each fitting model was computed. Figure 7 shows the median test deviance plus and minus one standard error over 10 simulations, plotted against the number of genes in each model.

Figure 8 shows the number of genes correctly included in the model versus the number of genes in the model, for each of CUS and lasso. Neither method is very good at isolating the true non-zero predictors, but surprisingly the lasso

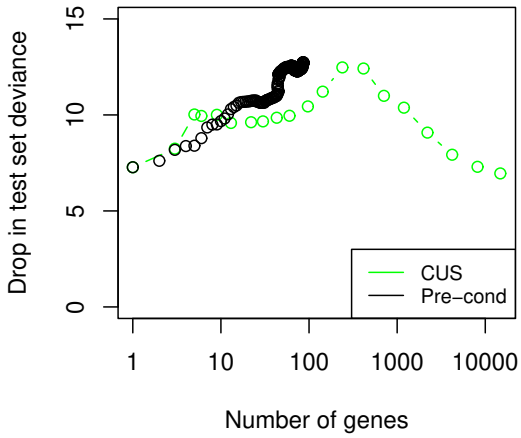

Figure 5: *Kidney cancer example: drop in test sample deviance for CUS (Cox univariate shrinkage), and pre-conditioning applied to CUS. In the latter we apply the regression version of the lasso to the fitted values from CUS.*

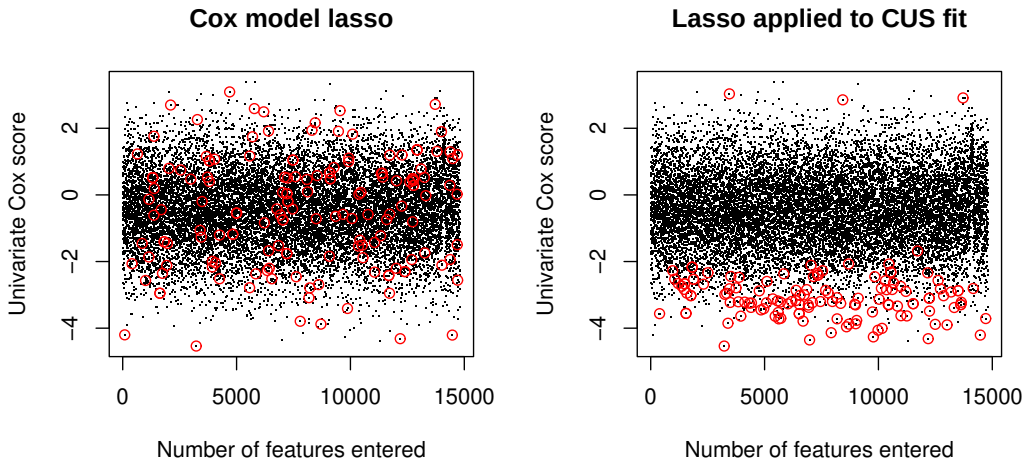

Figure 6: *Kidney cancer example: univariate Cox scores for all predictors (black) and those entered into the model (red). Panel on the left corresponds to L_1-penalized partial likelihood, fit by* `coxpath`*. On the right is the standard regression version of the lasso applied to the fitted values from CUS.*

DOI: 10.2202/1544-6115.1438

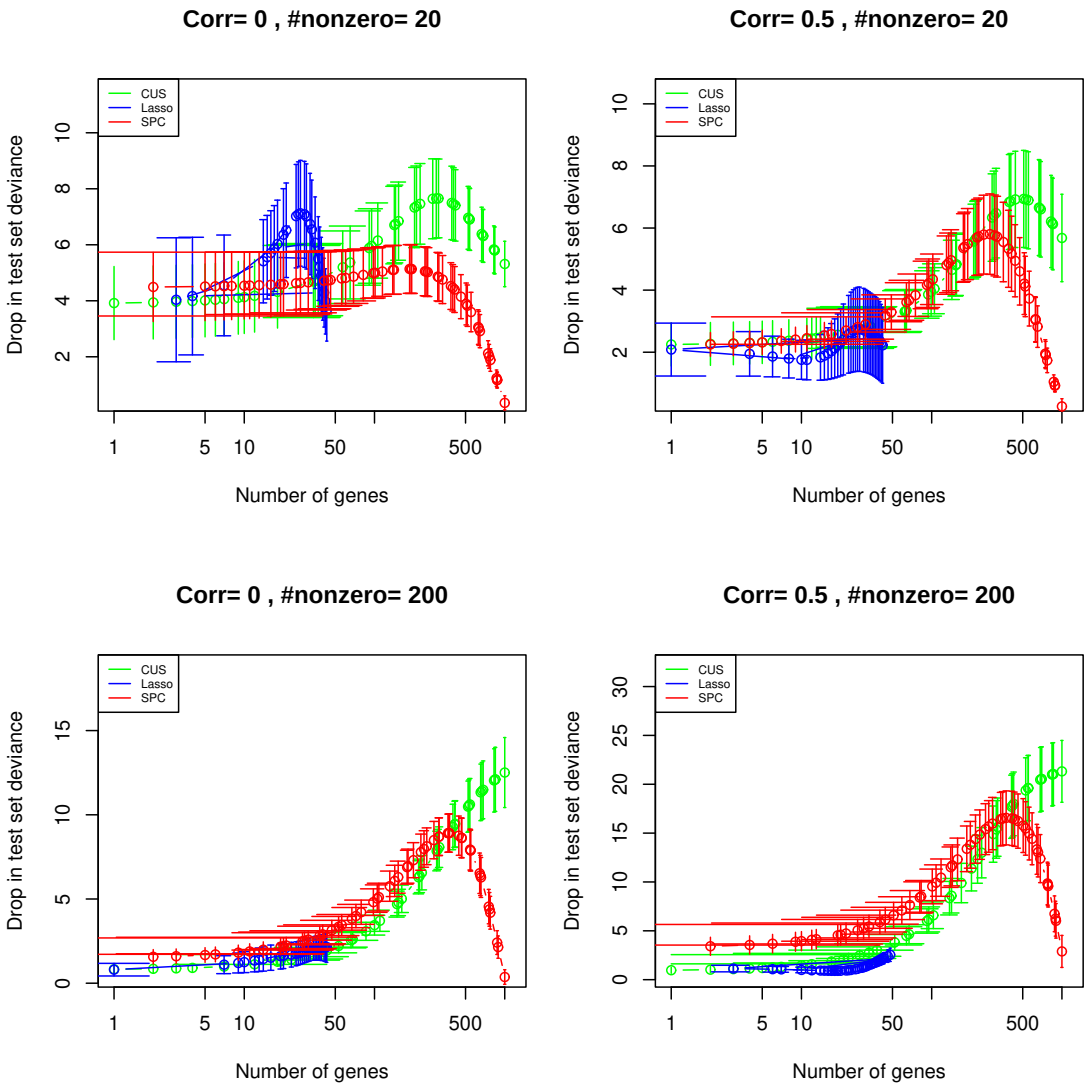

Figure 7: *Simulation results: average drop in test set deviance (± one standard error), for three different estimators, over four different simulation settings.*

does no better than the simple CUS procedure.

In Figures 9 and 10 we chose the model complexity by cross-validation and report the drop in test set deviance (Figure 9) and the C-index (Figure 10). The latter index is time-integrated area under the ROC curve, appropriate for survival studies (Heagerty & Zheng 2005). In every case the CUS performs on average as well or better than the lasso and supervised principal components).

5 Some real data examples

We applied three methods- lasso, supervised principal components, and CUS, to three real datasets. The first is the kidney data discussed earlier; the other datasets are the diffuse large cell lymphoma data of Rosenwald et al. (2002) (7399 genes and 240 patients), and the breast cancer data of van't Veer et al. (2002) (24881 genes, 295 patients). We randomly divided each dataset into training and test sets ten times, and averaged the results. The first two datasets included test sets of size 88 and 80 respectively, so we retained those test set sizes. For the third dataset, we took a test set of about one-third (98 patients), and to ease the computation load, restricted attention to the 10,000 genes with largest variance. Figure 11 shows the median test deviance plus and minus one standard error over 10 simulations, plotted against the number of genes in each model. The CUS method yields a larger drop in test deviance than the lasso, in at least two of the datasets and performs a little better than SPC overall.

The R package `uniCox` implementing the methods of this paper will soon be available on CRAN.

Appendix

Monotonicity of profiles. The subgradient equation for each β_j is:

$$\sum_{k=1}^{K}\Big(x_{kj} - d_k\frac{\sum_{m\in R_k} x_{mj}\exp(x_{mj}\beta_j)}{\sum_{m\in R_k}\exp(x_{mj}\beta_j)}\Big) - \lambda\cdot t_j = 0 \tag{12}$$

where $t_j \in \text{sign}(\beta_j)$, that is $t_j = \text{sign}(\beta_j)$ if $\beta_j \neq 0$ and $t_j \in [-1,1]$ if $\beta_j = 0$

Claim: $|\hat\beta_j(\lambda)|$ is strictly decreasing in λ when $\hat\beta_j(\lambda)$ is non-zero, and if $\hat\beta_j(\lambda) = 0$ then $\hat\beta_j(\lambda') = 0$ for all $\lambda' > \lambda$.

Proof: for ease of notation, suppose we have a single β having the subgradient equation

DOI: 10.2202/1544-6115.1438

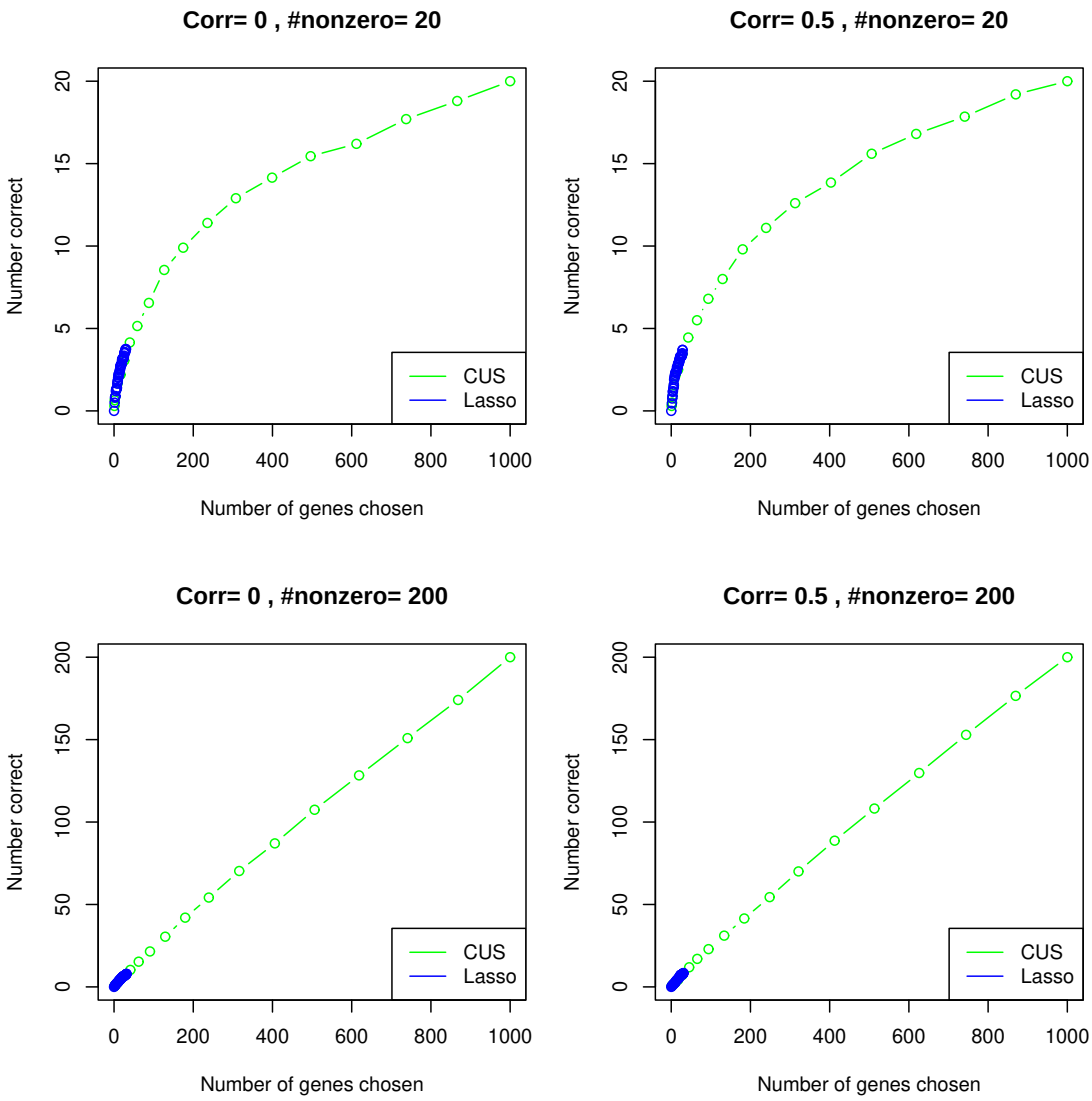

Figure 8: *Simulation results: number of correctly included predictors in the model, for CUS and lasso, over the four different simulation settings.*

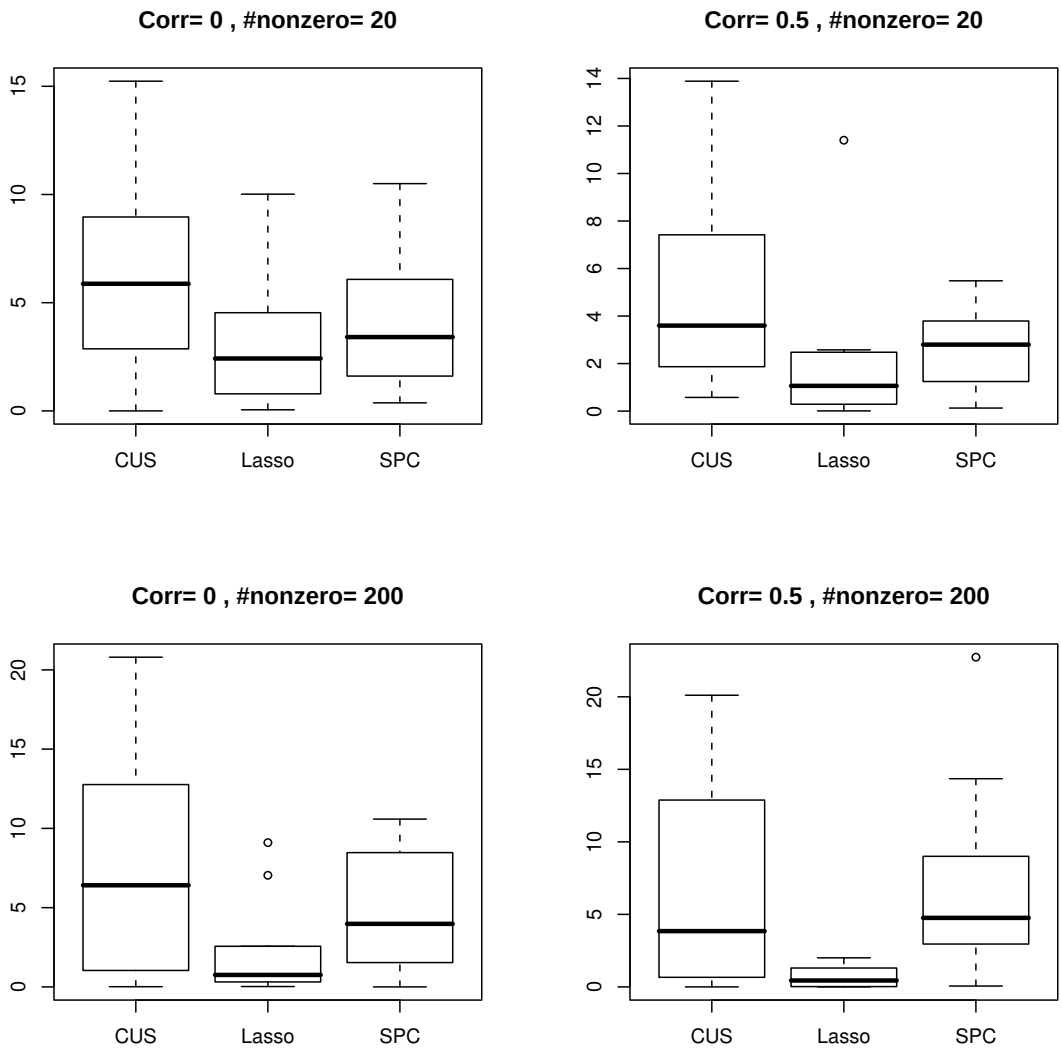

Figure 9: *Simulation results: average drop in test set deviance standard error), for three different estimators, over four different simulation settings.* The model tuning parameter was chosen in each case by cross-validation.

DOI: 10.2202/1544-6115.1438

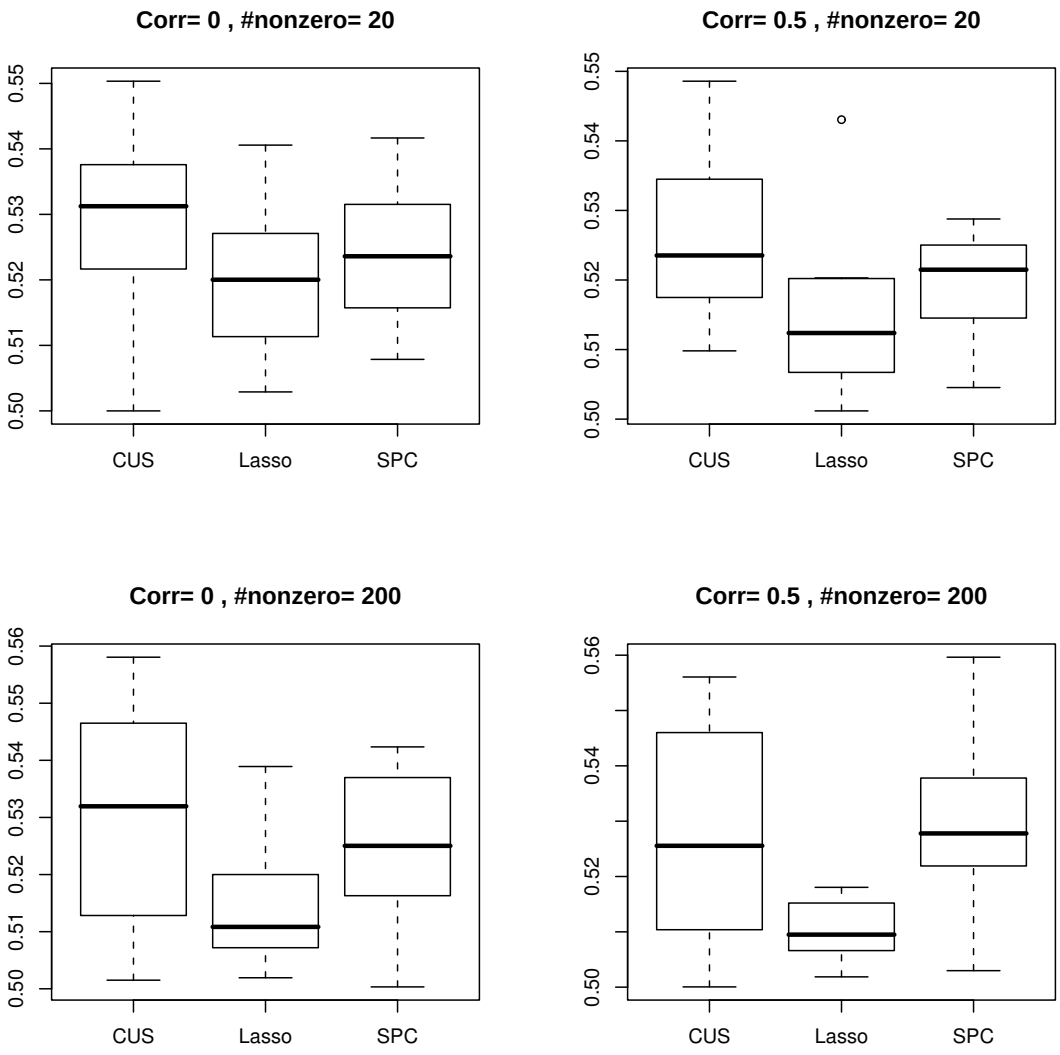

Figure 10: *Simulation results: average test-set C-index for three different estimators, over four different simulation settings. The model tuning parameter was chosen in each case by cross-validation.*

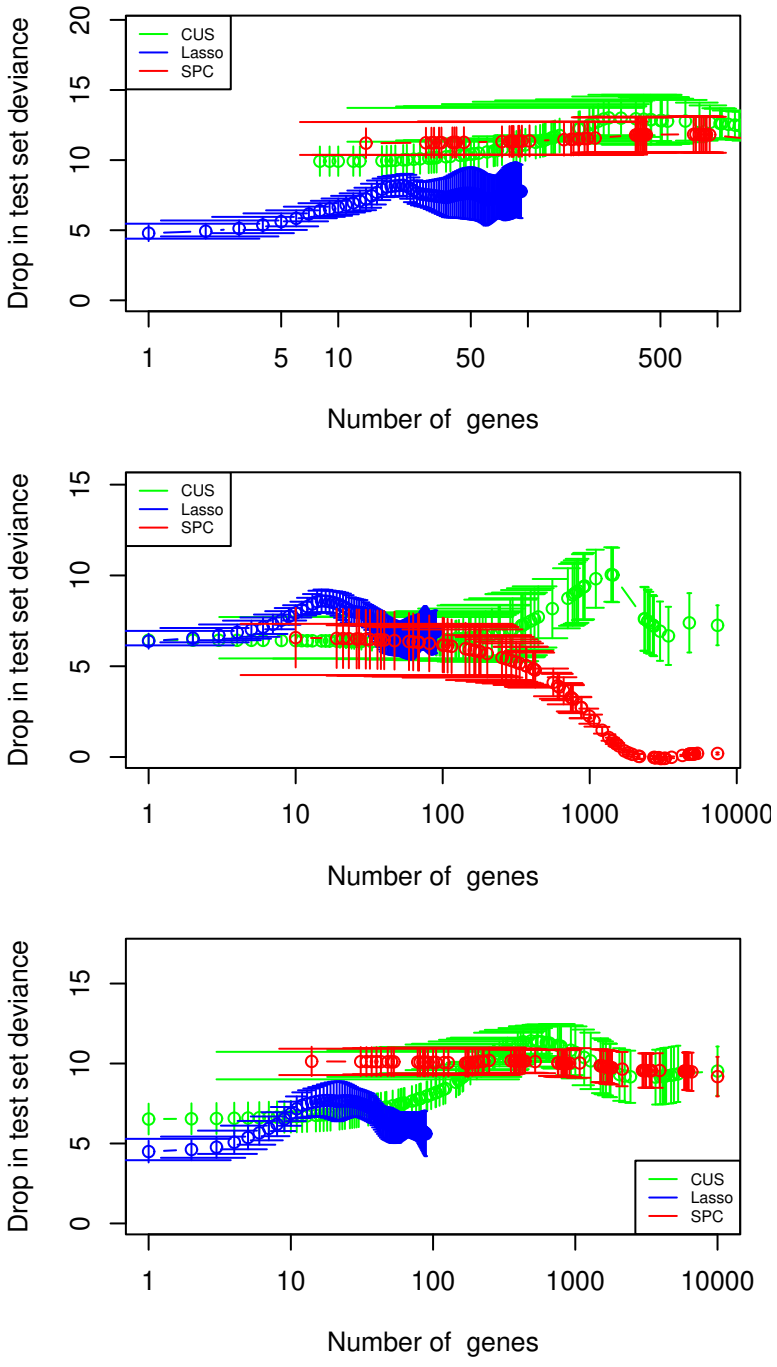

Figure 11: *Top to bottom: results for Kidney, Lymphoma and Breast cancer datasets; shown is the average drop in test set deviance (± one standard error) for CUS (Cox univariate shrinkage), lasso and SPC (supervised principal components).*

DOI: 10.2202/1544-6115.1438

$$g(\beta) - \lambda t = 0 \tag{13}$$

with $t \in \text{sign}(\beta)$.

Then $g'(\beta)$ is the negative of the observed information and it is easy to show that $g'(\beta) < 0$.

Denote the solution to (13) for parameter λ by $\hat{\beta}(\lambda)$. Suppose that we have a solution $\hat{\beta}(\lambda) \neq 0$ for some λ. WLOG assume $\hat{\beta}(\lambda) > 0$. Then by (13) we have

$$g(\hat{\beta}(\lambda)) = \lambda$$

Then if $\lambda' > \lambda$, we can't have $\hat{\beta}(\lambda') \geq \hat{\beta}(\lambda)$ since this would imply $g(\hat{\beta}(\lambda')) < g(\hat{\beta}(\lambda)) = \lambda < \lambda'$. Hence $\hat{\beta}(\lambda') < \hat{\beta}(\lambda)$.

On the other hand, if $\hat{\beta}(\lambda) = 0$ then

$$g(0) - \lambda t = 0$$

for some t with $|t| \leq 1$. Then if $\lambda' > \lambda$ the equation $g(0) - \lambda' t' = 0$ can be solved by taking $t' = t(\lambda/\lambda')$. Hence $\hat{\beta}(\lambda') = 0$.

Proof of (11). Let

$$g_j(\beta_j) = \sum_{k=1}^{K} \Big(x_{kj} - d_k \frac{\sum_{m \in R_k} x_{mj} \exp(x_{mj}\beta_j)}{\sum_{m \in R_k} \exp(x_{mj}\beta_j)} \Big)$$

Suppose $\hat{\beta}_j(0) > 0$. Then $\hat{\beta}_j(\lambda) \neq 0 \iff$ the equation $g_j(\beta_j) - \lambda = 0$ has a solution in $(0, \hat{\beta}_j(0))$. Recall that the x_j have been standardized by $\sqrt{V_j}$. Hence if the gradient for β_j for the raw (unnormalized) data is $U_j(\beta_j)$, with $U_j(0) \equiv U_j$, then $g_j(\beta_j) = U_j(\beta_j)/\sqrt{V_j}$ and we must have $U_j/\sqrt{V_j} > \lambda$. The quantity $U_j/\sqrt{V_j}$ is the square root of the usual score statistic for testing $\beta_j = 0$.

References

Bickel, P. & Levina, E. (2004), 'Some theory for Fisher's linear discriminant function,"naive Bayes", and some alternatives when there are many more variables than observations', *Bernoulli* **10**, 989–1010.

Efron, B. (2008), Empirical bayes estimates for large scale prediction problems. Unpublished.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultra-high dimensional feature space', *Journal of the Royal Statistical Society Series B, to appear* .

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

Paul, D., Bair, E., Hastie, T. & Tibshirani, R. (2008), '"pre-conditioning" for feature selection and regression in high-dimensional problems', *Annals of Statistics* **36**(4), 1595–1618.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. & Staudt, L. M. (2002), 'The use of molecular profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma', *The New England Journal of Medicine* **346**, 1937–1947.

Tibshirani, R. (1997), 'The lasso method for variable selection in the cox model', *Statistics in Medicine* **16**, 385–395.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2001), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proceedings of the National Academy of Sciences* **99**, 6567–6572.

Tusher, V., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.

van't Veer, L. J., van de Vijver, H. D. M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Anke T. Witteveen and, G. J. S., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Verweij, P. & van Houwelingen, H. (1993), 'Cross-validation in survival analysis', *Statistics in Medicine* **12**, 2305–2314.

Zhao, H., Tibshirani, R. & Brooks, J. (2005), 'Gene expression profiling predicts survival in conventional renal cell carcinoma', *PLOS Medicine* pp. 511–533.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society Series B.* **67**(2), 301–320.

DOI: 10.2202/1544-6115.1438