BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 16 2007, pages 2080–2087
doi:10.1093/bioinformatics/btm305

Gene expression

Predicting survival from microarray data—a comparative study

H.M. Bøvelstad[1,*,†], S. Nygård[1,†], H.L. Størvold[2], M. Aldrin[3], Ø. Borgan[1], A. Frigessi[4] and O.C. Lingjærde[2]

[1]Department of Mathematics, [2]Department of Informatics, University of Oslo, [3]Norwegian Computing Center and [4]Institute of Basic Medical Sciences, Department of Biostatistics, University of Oslo and Statistics for Innovation – (sfi)[2], Norway

Received on October 31, 2006; revised on May 24, 2007; accepted on May 28, 2007

Advance Access publication June 6, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Survival prediction from gene expression data and other high-dimensional genomic data has been subject to much research during the last years. These kinds of data are associated with the methodological problem of having many more gene expression values than individuals. In addition, the responses are censored survival times. Most of the proposed methods handle this by using Cox's proportional hazards model and obtain parameter estimates by some dimension reduction or parameter shrinkage estimation technique. Using three well-known microarray gene expression data sets, we compare the prediction performance of seven such methods: univariate selection, forward stepwise selection, principal components regression (PCR), supervised principal components regression, partial least squares regression (PLS), ridge regression and the lasso.

Results: Statistical learning from subsets should be repeated several times in order to get a fair comparison between methods. Methods using coefficient shrinkage or linear combinations of the gene expression values have much better performance than the simple variable selection methods. For our data sets, ridge regression has the overall best performance.

Availability: Matlab and R code for the prediction methods are available at http://www.med.uio.no/imb/stat/bmms/software/microsurv/.

Contact: hegembo@math.uio.no

1 INTRODUCTION

Prediction of cancer patient survival based on gene expression profiles is an important application of genome-wide expression data (e.g. Rosenwald *et al.*, 2002; van de Vijver *et al.*, 2002; van't Veer *et al.*, 2002). By uncovering the relationship between time to death (or time to another disease related adverse event) and the tumor expression profile, one hopes to achieve more accurate prognoses and improved treatment strategies. A substantial challenge in this context comes from the fact that the number of genomic variables p is usually much larger than the number of subjects n (i.e. $p \gg n$). It is very difficult to select the most powerful genomic variables for prediction, as these may depend on each other in an unknown fashion. Because of the large number of expression values, it is easy to find predictors that perform excellently on the fitted data, but fail in external validation, leading to poor prediction rules.

We study the situation where the response of interest is a possibly censored survival time, and the Cox proportional hazards model (Cox, 1972) is used for inference. Here, the instantaneous risk of an event at time t for an individual with gene expression values $\mathbf{x} = (x_1, \ldots, x_p)'$ is given by

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}), \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression parameters, and $h_0(t)$ is the baseline hazard. In the classical setting with $n > p$, the regression parameters are estimated by maximizing Cox's partial likelihood, but this does not work when $p > n$. Early papers proposing prediction rules based on expression data most often apply simple dimension reduction methods, like univariate selection rules. More sophisticated methods, adapting ideas from dimension reduction and penalized regression, have been proposed (e.g. Bair and Tibshirani, 2004; Bair *et al.*, 2006; Segal, 2006; van Houwelingen *et al.*, 2006). However, there is a lack of literature on validation and systematic comparison of prediction methods (Michiels *et al.*, 2005). Furthermore, the importance of considering several splits of the data into training and test sets when comparing methods has received little attention. Neglecting to do this may lead to biased results concerning the relative performance of methods.

The purpose of this article is 2-fold: first, we propose a systematic approach for comparison of methods that predict survival based on high-dimensional genomic data; second, we apply this approach to compare the prediction performance of seven methods that combine the Cox regression model with some method for dimension reduction or shrinkage. The seven methods are univariate and forward stepwise selection, that pick out subsets of the genes exhibiting strong effect on survival, two principal components-based methods, one method that uses partial least squares to summarize the gene expressions into a few linear combinations, and finally two methods that shrink parameter estimates towards zero by imposing a penalty on the partial likelihood.

Most methods for dimension reduction or shrinkage require the selection of a tuning parameter that determines the amount

*To whom correspondence should be addressed.

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

of dimension reduction or shrinkage. This applies to the seven studied methods as well; here the tuning parameter is the number of genes for univariate and forward stepwise selection, the number of linear combinations for principal components and partial least squares and the amount of shrinkage for penalized regression. While tuning parameters are sometimes chosen subjectively after careful consideration of results, the use of an automatic selection criterion is important in our context. This ensures a fair comparison between methods and also that the methods are tuned to predict well on novel data rather than on the training data themselves. In this article, all tuning parameters are determined by partial likelihood cross-validation (Verweij and van Houwelingen, 1993).

In assessing the performance of a prediction rule, we need to train and test the method on separate data. Failure to do so would favor methods that fit well to the specific data at hand rather than to novel data from the same population. We also need to take into account the variation in prediction performance that would result from choosing a different split of the data into a training data set and a test data set. More generally, performance assessment for the prediction rules must account for all model selections and training decisions made.

In this article, we compare the various strategies on three well known data sets of which two are on breast cancer and one is on diffuse large-B-cell lymphoma, see Chang *et al*. (2005), Rosenwald *et al*. (2002), Sørlie *et al*. (2003), van de Vijver *et al*. (2002), van Houwelingen *et al*. (2006), van't Veer *et al*. (2002). We find that the two simple selection rules perform poorly compared to the more advanced prediction methods applying principal components, partial least squares or penalized regression. In particular, the penalized regression method ridge regression shows the best performance in all our three data sets. These results are not substantially affected for any of the data sets by a 25 or 50% reduction of the training set sample size.

2 METHODS

We will compare the prediction performance of seven prediction methods for Cox's proportional hazards model. In this section, we first review the Cox log partial likelihood and describe the seven methods. For all methods, a tuning parameter needs to be decided. This will be done using partial likelihood cross-validation, described in Section 2.3.

2.1 The Cox partial likelihood

When dealing with survival data and Cox's proportional hazards model in the classical setting with $n>p$, estimation of the regression coefficients $\boldsymbol{\beta}$ can be done by maximizing the log partial likelihood:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_i'\boldsymbol{\beta} - \log\left[\sum_{j \in R(t_i)} \exp(\mathbf{x}_j'\boldsymbol{\beta}) \right] \right\}. \quad (2)$$

Here t_i is the observed or censored survival time for individual i, and δ_i is an indicator for whether the survival time is observed ($\delta_i = 1$) or censored ($\delta_i = 0$). Further, $R(t_i)$ is the risk set at time t_i, i.e. the set of all individuals who are still under study just prior to time t_i, and $\mathbf{x}_i$ is the vector of normalized gene expression values for individual i.

2.2 Methods for prediction

2.2.1 Univariate selection For the univariate selection method, we first test the effect each gene expression value has by itself on survival. This can be done by adopting a univariate Cox model $h(t|x_g) = h_0(t)\exp(\beta_g x_g)$ for each gene g, and then test the null hypothesis $\beta_g = 0$ versus the alternative $\beta_g \neq 0$ using the score test (e.g. Klein and Moeschberger, 2003, Chapter 8.3). After testing the genes one-by-one, we arrange them according to increasing P-values. We then pick out the λ top ranked genes, and include them in a multivariate Cox regression model (1). Here, λ is a tuning parameter representing the number of genes to include in the model. We use the score test since it does not require the estimated regression coefficients. When dealing with large data sets, this feature contributes to a considerable reduction of computational time compared to the likelihood ratio and Wald tests.

2.2.2 Forward stepwise selection As opposed to the univariate selection method, the forward stepwise selection method takes into account possible correlation between the gene expressions. First, the most significant gene is found and ranked as number one using the univariate score test as described above. Then, all models with this gene and one more gene are tested versus the model with only the first gene. These tests are performed using the local score test (e.g. Klein and Moeschberger, 2003, Chapter 8.5). The new gene from the most significant model is selected, ranked as number two and included in a multivariate Cox regression model (1) along with the first ranked gene. We continue to rank genes in this manner and include them one-by-one until we have selected λ genes. Also here the use of a (local) score test gives a computational advantage compared to the likelihood ratio and Wald tests.

2.2.3 Principal components regression Principal components analysis exploits that gene expressions might be highly correlated, summarizing these by a single linear combination. After using principal components analysis to decompose the set of p gene expressions, we pick out the λ first principal components, i.e. the first λ linear combinations that account for as much variation in the gene expressions as possible. Then we include these principal components as covariates in a Cox regression model yielding principal components regression (PCR), see, e.g. Hastie *et al*. (2001, Chapter 3.4.4) for a review of PCR for the classical case of linear models. In principal components analysis the survival times are not used when forming and selecting the principal components, so no special theory is needed to perform PCR for censored survival data.

2.2.4 Supervised principal components regression A possible draw-back of PCR is that we have no guarantee that the selected principal components are associated with patient survival. That is, directions with high variability in the gene expressions can be due to effects not related to survival. With this argument in mind, Bair and Tibshirani (2004) and Bair *et al*., (2006) proposed the supervised principal components regression. This procedure first picks out a subset of the gene expressions that is correlated with survival by using univariate selection, and then applies PCR to this subset. In our analysis, we will pick out λ_1 percent of the top ranked genes according to the P-values obtained using univariate selection. Then, we apply principal components analysis to this subset of genes and include λ_2 of the first principal components into a multivariate Cox model (1) yielding PCR. For this method the tuning parameter $\lambda = (\lambda_1, \lambda_2)$ is bivariate.

2.2.5 Partial least squares regression Ordinary partial least squares (PLS) regression for linear models performs regression of the outcome on a smaller number λ of components which are linear combinations of the original covariates (e.g. Martens and Næs, 1989).

It can be shown that these components are the ones that maximize the covariance with the outcome. Since the ordinary PLS algorithm assumes a linear relation between the outcome and the covariates, this PLS method is not directly applicable to the non-linear Cox model. We will adopt the approach of Park *et al.* (2002) in which the full likelihood for Cox's model is reformulated as the likelihood of a Poisson model, i.e. a generalized linear model (GLM). This reformulation enables application of the iteratively reweighted partial least squares (IRPLS) procedure for GLM (Marx, 1996). In the implementation of the PLS algorithm of Park *et al.* (2002) the PLS components become a mixture of gene expression values and the baseline hazard. By treating the latter as an offset in the PLS regression, we obtain PLS components that depend solely on the gene expressions. See Nygård *et al.* (2006) for a detailed description of our PLS Cox method.

2.2.6 Ridge regression Ridge regression (Hoerl and Kennard, 1970) shrinks the regression coefficients by imposing a penalty on their squared values. For the Cox regression setting, the regression coefficients are estimated by maximizing the penalized log partial likelihood $l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \beta_j^2$, where $l(\boldsymbol{\beta})$ is the log partial likelihood given in (2) and $\lambda \sum_{j=1}^{p} \beta_j^2$ is the penalty term (Verweij and van Houwelingen, 1994). Here λ is the tuning parameter controlling the amount of shrinkage.

2.2.7 Lasso Like ridge regression, the lasso (Tibshirani, 1996) shrinks the regression coefficients toward zero by penalizing the size of the coefficients, but this time with the absolute values instead of the squared values. The penalized log partial likelihood thus becomes $l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|$ (Tibshirani, 1997). As before, $l(\boldsymbol{\beta})$ is the log partial likelihood (2), and λ is the tuning parameter determining the amount of shrinkage. Penalizing with the absolute values has the effect that a number of the estimated coefficients will become exactly zero, which means that the lasso also is a variable selection method. The lasso is a computationally demanding method. In our study, we have applied the lasso implementation of Cox regression due to Park and Hastie (2006), available through the R package glmpath. Another option could have been the proposal of Segal (2006).

2.3 Choosing the tuning parameter by cross-validation

The model complexity of the prediction methods is decided by estimating a tuning parameter. This estimation can be done in a number of ways; the most common being cross-validation. We will use the cross-validation criterion proposed by Verweij and van Houwelingen (1993) which is based on the Cox log partial likelihood (2).

The general idea of cross-validation is that one partitions the n individuals under study into K different folds, yielding K-fold cross-validation ($1 < K \leq n$). For a given value of the tuning parameter λ the prediction model is estimated using $K-1$ folds, and the adequacy of the model is assessed by applying the estimated model to the individuals in the left-out fold. This procedure is repeated K times, leaving one fold out at a time. The K estimates of prediction capability are then combined, and this serves as a measure of prediction capability for the model. This procedure is repeated for a range of λ values, and the chosen tuning parameter is the one optimizing the prediction capability.

A complication when cross-validation is applied to Cox's model (1), is that the terms in the Cox log partial likelihood (2) are not independent. Verweij and van Houwelingen (1993) introduced a leave-one-out ($K = n$) cross-validation criterion that circumvents this problem. For linear models the use of K-fold ($K < n$) cross-validation has proved to be favorable to the leave-one-out method (Shao, 1993), and we expect this to be the case for Cox regression as well. We will therefore use a more general form of the cross-validation criterion of Verweij and van Houwelingen (1993) that enables K-fold cross-validation. Also, when dealing with data of a considerable size, this generalization is important to reduce computational time. We will use $K = 10$ yielding 10-fold cross-validation.

Let $l(\boldsymbol{\beta})$ denote the Cox log partial likelihood (2), and let $l_{(-k)}(\boldsymbol{\beta})$ be the log partial likelihood when the kth fold is left out; $k = 1, \ldots, K$. Further denote by $\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)$ the estimate of $\boldsymbol{\beta}$ that is obtained by a given prediction method when the kth fold is left out and λ is the tuning parameter. Then the K-fold cross-validated log partial likelihood is given by

$$CV(\lambda) = \sum_{k=1}^{K} \{l(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)) - l_{(-k)}(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda))\}. \quad (3)$$

The optimal tuning parameter λ is obtained by maximizing $CV(\lambda)$. In practice, we maximize $CV(\lambda)$ using either a range of natural numbers or a fine grid of values to represent λ. We use the former for univariate selection, forward stepwise selection, PCR and PLS regression, and the latter for ridge regression and the lasso. For the supervised PCR the tuning parameter $\lambda = (\lambda_1, \lambda_2)$ is bivariate, and we use a combination of the two. That is, λ_1 and λ_2 are found by maximizing over a fine grid of cutoff P-values and a range of natural numbers, respectively.

3 STRATEGY FOR COMPARING METHODS

In this section, we describe how we compare the performance of the prediction methods of Section 2. The comparison will be done in terms of three different model evaluation criteria applied to the data sets mentioned in the introduction. Let us have a brief visit of the data sets before we describe the criteria in more detail.

3.1 Data sets

We will apply all methods to three data sets containing large-scale microarray gene expression measurements from tumor biopsies of cancer patients together with their (possibly censored) survival times.

3.1.1 Dutch breast cancer data The first data set is from van Houwelingen *et al.* (2006), and consists of $p = 4919$ gene expression measurements from $N = 295$ women with breast cancer. The data set is a subset of the data with 24 885 gene expression values from van de Vijver *et al.* (2002), which again is an extension of the data set with $N = 117$ patients from van't Veer *et al.* (2002).

3.1.2 DLBCL data This data set is from Rosenwald *et al.* (2002) and contains $p = 7399$ gene expression measurements from $N = 240$ patients with diffuse large-B-cell lymphoma (DLBCL).

3.1.3 Norway/Stanford breast cancer data The last data set is from Sørlie *et al.* (2003) and contains gene expression measurements from $N = 115$ women with breast cancer. We used the list of $p = 549$ intrinsic genes introduced in Sørlie *et al.* (2003). Missing values were imputed using 10-nearest neighbor imputation.

3.2 Model evaluation criteria

In order to evaluate the methods we first divide each data set randomly into two parts; one *training set* of about 2/3 of the patients used for estimation and one *test set* of about 1/3 of the patients used for evaluation or testing of the prediction capability of the estimated model. The description of the model evaluation criteria will be illustrated using one such 2:1 split of the Dutch breast cancer data.

Let us denote the parameter estimate from the training data for a given method by $\hat{\boldsymbol{\beta}}_{\text{train}}$. This estimate is found in two steps as described in Section 2.3 using 10-fold cross-validation: we first use 10-fold cross-validation to find the optimal tuning parameter value, $\hat{\lambda}_{\text{train}}$, and then perform estimation on the whole training set using $\hat{\lambda}_{\text{train}}$ to obtain $\hat{\boldsymbol{\beta}}_{\text{train}}$.

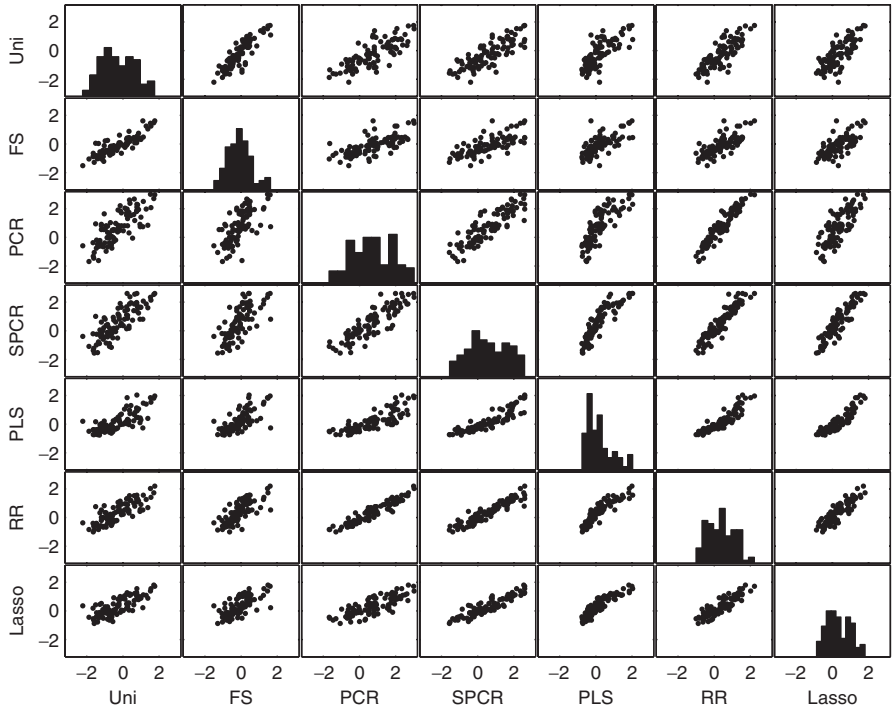

Fig. 1. Scatter plot matrix with histogram of prognostic indices (PI) on the diagonal and pairwise comparisons of the PIs using the seven different prediction methods on the off-diagonal for the first random training/test split of the Dutch breast cancer data. Uni = Univariate selection, FS = Forward stepwise selection, PCR = Principal component regression, SPCR = Supervised PCR, PLS = Partial least squares regression and RR = Ridge regression.

For each patient i in the test set, this estimate is then used together with its vector of gene expression values $\boldsymbol{x}_i$ to derive a prognostic index (PI) for the patient, given by

$$\hat{\eta}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\text{train}}.$$

The scatter plot matrix in Figure 1 shows on the diagonal histograms of the PIs for each of the methods and on the off-diagonal pairwise scatter plots of the PIs calculated on our 2:1 training/test split of the Dutch breast cancer data. From the scatter plot matrix we see that the PIs from the different methods are highly correlated, but still being quite far from laying on straight lines, meaning that the PIs from the methods can be quite different.

The performance of a method is good when the PIs explain the actual survival times of the test patients. We will now describe three different criteria for evaluating how well the survival times are explained. For all criteria we also compare with the null model, i.e. the model where all the PIs are equal to zero. This corresponds to a model that does not use any gene expression data for prediction, and hence gives the same prediction of survival for all patients in the test data set. It is worth noting that a method can perform *worse* than the null model. This has to do with the concept of overfitting and the fact that an overfitted model may predict poorer than a prediction rule that gives the same prediction for all patients.

3.2.1 Log rank test In cancer treatment one is often interested in assigning the patients to subgroups based on their prognosis, e.g. into one with 'good' and one with 'bad' prognosis. We will mimic this approach using equally sized subgroups. Thus, a patient i in the test set is assigned to the 'bad' group if its prognostic index $\hat{\eta}_i$ is above the median of the prognostic indices. In order to evaluate the performance of this grouping, we apply a log rank test and use the P-value as an evaluation criterion. The P-values obtained for all seven prediction methods from our split of the Dutch breast cancer data are shown in the left panel of Table 1. A disadvantage with this criterion is that it just evaluates whether the patients are assigned to the 'right group', and not how well the patients are ranked within the groups. Moreover, the underlying pathology is more likely to be 'continuous' rather than categorical, making a grouping less reasonable. Finally, if there exist

Table 1. Values of the three model evaluation criteria for the various prediction methods using two random training/test splits of the Dutch breast cancer data

Method	Split 1			Split 2		
	LR	PI	Dev	LR	PI	Dev
Uni	0.0012	0.0149	−1.4	0.5	0.5	0.0
FS	0.0064	0.0237	−3.5	0.5	0.5	0.0
PCR	0.0002	0.0008	−7.8	0.0003	0.0014	−10.1
SPCR	0.0704	0.0161	1.9	0.0000	0.0000	−16.1
PLS	0.0040	0.0291	−1.6	0.0003	0.0005	−11.4
RR	0.0010	0.0015	−9.8	0.0020	0.0000	−17.9
Lasso	0.0481	0.0094	−6.4	0.0018	0.0004	−11.5

The criteria are based on the log rank test for two groups (LR), Cox regression on prognostic index (PI) for and difference in deviance from the null model (Dev). The P-values of 0.5 and deviance values of 0 for the first two methods when applied to split 2 indicate that no genes were selected. For method abbreviations, see the legend of Figure 1.

dichotomous subgroups of the disease, a classification according to the median PI may not be biologically meaningful. One option to overcome the latter problem would have been to use the area under the ROC curve to evaluate how well the PIs are able to classify the patients into 'good' and 'bad' prognosis groups (Segal, 2006).

3.2.2 Prognostic index This approach is trying to overcome the drawbacks of the first criterion by using the predictions individually instead of group-wise. Here, we use the prognostic index $\hat{\eta}_i$ as a single continuous covariate in a Cox regression on the test data set, i.e. we fit the model

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\hat{\eta}_i\alpha). \quad (4)$$

We then test the null hypothesis $\alpha = 0$ versus the alternative $\alpha \neq 0$ using the likelihood ratio test, and again look at the *P*-value to evaluate a method's performance. The *P*-values from our training/test split of the Dutch breast cancer data are shown in the left panel of Table 1.

3.2.3 Deviance The third measure we will use to evaluate the predictions is based on the deviance. We use the difference in deviance between a fitted model and the null model, given by

$$-2\{l^{(\text{test})}(\hat{\boldsymbol{\beta}}_{\text{train}}) - l^{(\text{test})}(\mathbf{0})\}. \quad (5)$$

Here $l^{(\text{test})}(\hat{\boldsymbol{\beta}}_{\text{train}})$ and $l^{(\text{test})}(\mathbf{0})$ are the Cox log partial likelihood (2) for the test data evaluated at $\hat{\boldsymbol{\beta}}_{\text{train}}$ and $\mathbf{0}$, respectively. The performance is good when (5) is small. The values of this criterion calculated for the seven prediction methods on the Dutch breast cancer data are given in the left panel of Table 1.

3.3 Multiple splits in training and test sets

From only one split of the data into training and test sets, we will not know to which extent the resulting criteria values depend on the actual training/test randomization. To illustrate the dependence on the split, we evaluate the prediction performance using the three criteria on another 2:1 training/test split of the Dutch breast cancer data. The results are shown in the right panel of Table 1. If we just look at our first split, PCR has the best performance according to the first two prediction evaluation criteria, whereas ridge regression has the best performance based on the deviance criterion. This is in contrast with the results from split 2, where PLS regression and supervised PCR have equal or better performance than PCR according to the log-rank test, while supervised PCR, PLS regression, ridge regression and the lasso perform better than PCR for the prognostic index criterion.

The inconsistency of the results from splits 1 and 2 illustrates that several splits are needed to get a reliable evaluation of the performance of the prediction methods. Another drawback of investigating only one split is that we do not fully take advantage of all the available information in the data. Therefore, we divide the data *S* times at random into 2:1 training/test sets and calculate the three criteria for each of these splits, yielding *S* different values for each of the three criteria. We then look at the median and the spread of the *S* values. We will use $S = 50$; a number seeking to strike a balance between uncertainty due to few splits on one hand and long computational time on the other. In these splits we made sure that there were at least one event in each fold. A summary of our procedure for evaluating the performance of the prediction methods is given in Table 2 for easy reference.

Table 2. Summary of the procedure for evaluating the performance of the prediction methods

- For each prediction method:
  - For each of *S* random splits into training and test data sets:
    - Find the optimal tuning parameter $\hat{\lambda}_{\text{train}}$ by *K*-fold cross-validation using the training data set, see Section 2.3.
    - Given $\hat{\lambda}_{\text{train}}$, estimate the vector of regression coefficients $\hat{\boldsymbol{\beta}}_{\text{train}}$ on the whole training data set.
    - Calculate the values of the three performance criteria on the test data set as described in Section 3.2.
- Compare the different prediction methods in terms of the median and the spread of the criteria values.

4 RESULTS

The results after applying the evaluation procedure described in Table 2 to each of the seven prediction methods for each of the three data sets described in Section 3.1 are summarized in the boxplot matrix of Figure 2. We will consider the median of the 50 values obtained from our three prediction performance criteria as the outcome of main interest. The rows of the boxplot matrix correspond to the three performance criteria, and the columns correspond to the three data sets. A horizontal line indicating the null model with no gene information included is displayed for reference. Note that the *P*-values of the log-rank test criterion and prognostic index criterion are given on the $\log_{10}$ scale, and that for these criteria the null model is represented by a *P*-value of 0.5.

Investigating Figure 2, it is clear that methods using univariate and forward stepwise selection have the poorest performance among the seven prediction methods. This is consistent over the three prediction performance criteria and over the three data sets. In fact, for the DLBCL data, the two methods pick out the null model (with no genes) as the best prediction model in more than 50% of the 50 splits. The prediction performance of the five other methods are more comparable, even though the results indicate a favor of ridge regression, at least in the first two data sets. We also note that for all the three data sets PCR has a better prediction capability than supervised PCR.

There is a fairly large spread of values in the boxplots of Figure 2 over the 50 splits. This is due to the variation caused by splitting the data at random into training and test sets as well as to the variation in the performance of the prediction methods for given splits. In order to get an impression of how much of the variation that is due to the prediction methods, we used ridge regression as a benchmark, and for each of the six other methods computed the difference between their deviance and the deviance of ridge regression. Figure 3 shows boxplots of these differences in deviance for the 50 splits. All the median values are positive, which shows that ridge regression is performing consistently better than the other methods in terms of the deviance criterion. The other two criteria showed similar results.

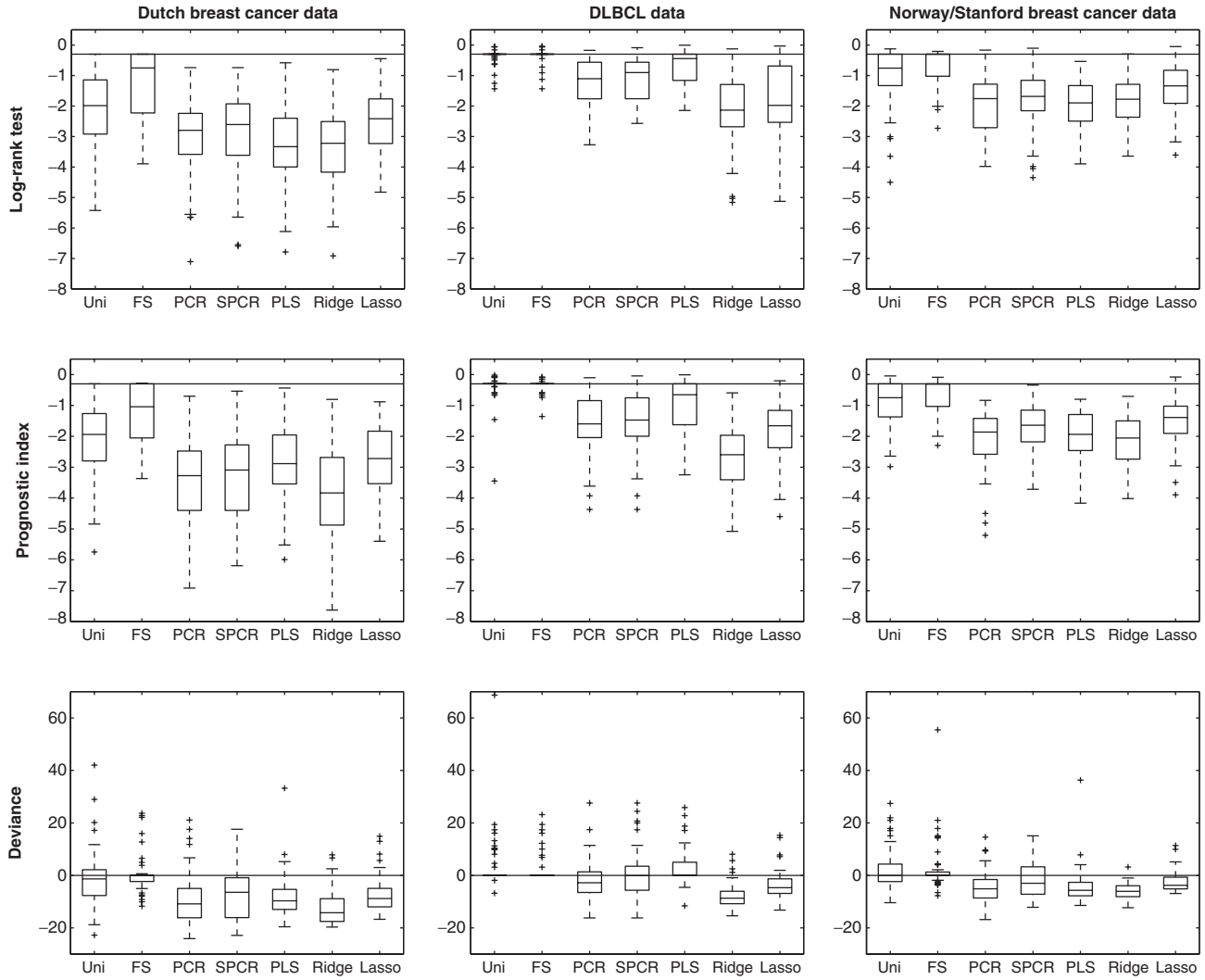

Fig. 2. The boxplot matrix gives the results after applying the seven prediction methods to each of 50 training/test splits of each of the three data sets. The rows of the boxplot matrix represent the three performance criteria, and the columns represent the three data sets. As a reference, the horizontal line in each plot represents the null model with no gene information included. Note that the *P*-values of both the log-rank test and the prognostic index are represented on the $\log_{10}$ scale. A small value of a criterion corresponds to a good prediction performance. Note that for the DLBCL data, both univariate and forward stepwise selection pick out the null model with no genes ($\lambda = 0$) in more than 50% of the 50 splits, and thus no box can be displayed. For method abbreviations, see the legend of Figure 1.

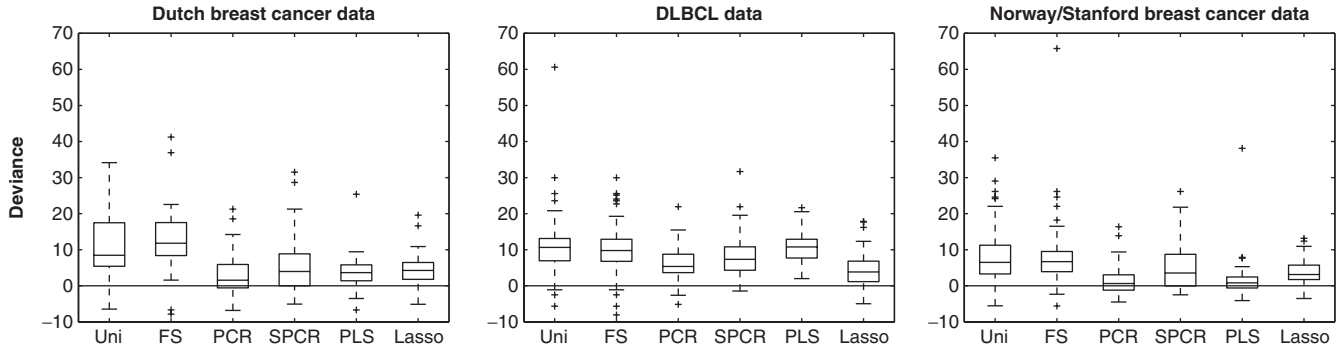

Fig. 3. The boxplots give the difference in deviance between the six methods and ridge regression, and thus illustrates the variation due to regression methods corrected for the variation due to the 50 random training/test splits. For method abbreviations, see the legend of Figure 1.

Table 3. The 0th, 25th, 50th, 75th, and 100th percentiles of the optimal tuning parameter $\hat{\lambda}$ obtained from estimation on 50 training/test splits of the three data sets

Method	Dutch breast cancer data					DLBCL data					Norway/Stanford breast cancer data				
	0	25	50	75	100	0	25	50	75	100	0	25	50	75	100
Uni	0	2	3	5	13	0	0	0	1	11	0	0	1	3	10
FS	0	0	1	1	3	0	0	0	0	2	0	0	1	1	3
PCR	1	7	8	14	24	0	4	9	11	16	1	2	3	5	9
SPCR															
Percentage of genes	0.1	1	5	50	100	0	2.5	100	100	100	0.25	2.5	25	100	100
Number of PCA directions	1	3	6	8	16	0	3	5	10	16	1	1	2	4	9
PLS	1	1	1	1	2	0	0	1	1	2	1	1	1	1	2
RR	100	200	200	200	400	1280	2560	2560	5120	5120	160	320	320	320	640
Lasso															
Tuning parameter	4.0	5.3	6.1	7.1	9.3	14.2	24.8	28.5	32.8	73.9	7.3	12.5	15.7	20.1	25.2
Number of genes	4	10	14	25	36	0	6	10	16	50	1	3	5	8	18

Since the tuning parameter for the supervised PCR is bivariate, both the percent of top rank genes and the number of PCA directions are given in the table. For the lasso we present both the optimal tuning parameter and the corresponding number of non-zero estimated regression coefficients. See Section 2.2 for description of the tuning parameters for the different methods. For method abbreviations, see the legend of Figure 1.

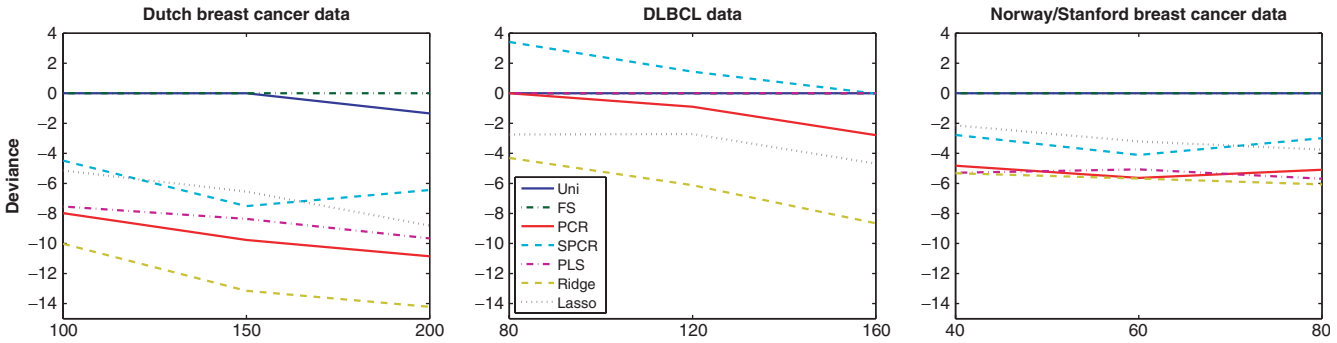

Fig. 4. Illustration of the prediction performance when reducing the size of the training set twice (25 and 50% compared to the original size). The reduction was done for all 50 splits into training and test set; the test set was put aside and kept unchanged during the whole procedure. This was done at random, with the constraint that each fold should contain at least one event. Only the median values of the deviance criterion are represented, and the values of the largest training sets correspond to the median values of the boxplots in the third row of Figure 2. For method abbreviations, see the legend of Figure 1.

In Table 3 the 0th, 25th, 50th, 75th, and 100th percentiles of the optimal tuning parameter $\hat{\lambda}$ obtained from estimation on the 50 training/test splits of the three data sets are given. We see that the optimal tuning parameter also vary quite a lot from split to split.

Despite the difference in sample size among the three data sets under study, the results seem to be rather similar. The univariate selection and the forward stepwise selection methods perform poorly, whereas the remaining five methods are more comparable, with ridge regression having the overall best prediction performance. We have also studied the performance of the methods when reducing the size of the training sets. For each of the three data sets, the number of individuals in each fold in the original training data set was reduced by 25 and 50% This was done at random, with the constraint that each fold should contain at least one event. The test set was kept unchanged during the whole procedure. The results obtained from this sample size reduction are displayed for the deviance criterion in Figure 4. From the figure, we observe that the differences between the methods are consistent over the range of training set sample sizes, sustaining the conclusions drawn from analyzing the full data sets. We also note that the expected weakening in prediction power due to the decrease in training set sample size is less clear in the Norway/Stanford breast cancer data than in the other two data sets. This may be an effect of the constraint of having at least one event in each fold.

5 DISCUSSION AND CONCLUSIONS

Several methods have recently been proposed to predict survival based on gene expression data, where the number of

gene expressions is much larger than the number of individuals (or biological replicates). To our knowledge, few papers have systematically compared the performance of the various methods used for predicting survival. To help potential users to choose an appropriate method for prediction, we have studied seven methods based on the Cox proportional hazards model applied to three well-known data sets.

All methods have a tuning parameter that must be chosen. To make a fair comparison among methods, we used the automatic cross-validation procedure of Verweij and van Houwelingen (1993) to perform that choice. Our study was based on splitting the data into a training set (used for estimating the model) and a test set (used to evaluate the prediction performance). We demonstrated that the results were sensitive to the split into training and test sets, and used therefore 50 random splits instead of only one.

The methods applying univariate and forward stepwise selection performed considerably worse than the five other methods we have studied. Overall, ridge regression tends to be the best method in our study. These conclusions are in accordance with what has been found for ordinary linear regression; see for instance Aldrin (1997) and Frank and Friedman (1993). Therefore, we strongly suspect that our conclusions would hold for most data sets with censored survival times, even though our present study is based on only three data sets. Interestingly, ordinary PCR performed slightly better than supervised PCR on all three data sets. This is in contrast with the results of Bair and Tibshirani (2004) and Bair *et al*. (2006), but here only one random split of the data into training and test sets was used.

In this article, we have focused on finding the best prediction rule for the time to an adverse event using all the available gene expression measurements in a microarray data set. In many studies, however, the main focus is on finding a small subset of the genes that are the most important ones for some phenotype. These genes are then often investigated further in order to obtain a better mechanistic understanding of the biological processes involved in the phenotype development. But performing subset selection can also be of interest for prediction tasks, for instance by picking out a few genes for simpler and cheaper diagnostic arrays. With these purposes in mind, we find the penalized regression method lasso very interesting, as it also is a variable selection method, but one with considerably better prediction performance than our other two, much simpler, selection methods. We note that the lasso picked out as few as 14, 10, and 5 genes (medians over the 50 splits, cf. Table 3) for the three data sets. These genes could well be represented on arrays from high-sensitivity platforms such as RT-PCR or even protein arrays.

ACKNOWLEDGEMENTS

The work of S.N. and M.A. was financed by the Norwegian Research Council (NFR) through Statistical Analysis of Risk (project number 154079/420), and the work of H.M.B. by NFR via Statistical Methodologies for Genomic Research (project number 167485). Part of the work of H.M.B., S.N. and Ø.B. was done during a research stay at the Center for Advanced Study, Oslo, Norway, over the academic year 2005/2006. The center is acknowledged for the excellent working facilities and inspiring environment provided there. Finally, the authors thank H. van Houwelingen and T. Sørlie for providing the two breast cancer data sets. Funding to pay the open access charges was provided by Statistics for Innovation – sfi^2.

Conflict of Interest: none declared.

REFERENCES

Aldrin,M. (1997) Length modified ridge regression. *Comput. Stat. Data Anal.*, **25**, 377–398.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, 511–522.

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

Chang,H.Y. *et al*. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA.*, **102**, 3738–3743.

Cox,D.R. (1972) Regression models and life tables (with discussion). *J. R. Stat. Soc. B*, **34**, 187–220.

Frank,I.E and Friedman,J.H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–147.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.

Klein,J.P. and Moeschberger,M.L. (2003) *Survival Analysis. Techniques for Censored and Truncated Data*. 2nd edn. Springer-Verlag, New York.

Martens,H. and Næs,T. (1989) *Multivariate Calibration*. Wiley, New York.

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

Michiels,S. *et al*. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Nygård,S. *et al*. (2006) Partial least squares Cox regression on genomic data handling additional covariates. Statistical Research Report 5/2006. Department of Mathematics, University of Oslo. http://www.math.uio.no/eprint/stat_report/2006/05-06.html

Park,M.P. and Hastie,T. (2006) L1 regularization path algorithm for generalized linear models. *Technical report*. 2006–14. Department of Statistics, Stanford University. http://www-stat.stanford.edu/reports/papers2006.html

Park,P.J. *et al*. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18** (Suppl. 1), 120–127.

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

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

Shao,J. (1993) Linear model selection by cross-validation. *J. Am. Stat. Assoc.*, **88**, 486–494.

Sørlie,T. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA.*, **100**, 8418–8423.

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

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

van de Vijver,M.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van Houwelingen,H.C. *et al*. (2006) Cross-validated Cox regression on micro-array gene expression data. *Stat. Med.*, **25**, 3201–3216.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Verweij,P.J.M. and van Houwelingen,H.C. (1993) Cross-validation in survival analysis. *Stat. Med.*, **12**, 2305–2314.

Verweij,P.J.M. and van Houwelingen,H.C. (1994) Penalized likelihood in Cox regression. *Stat. Med.*, **13**, 2427–2436.