

Constructing Multivariate Prognostic Gene Signatures with Censored Survival Data

Derick R. Peterson, PhD
Department of Biostatistics and Computational Biology
University of Rochester
Rochester, New York 14642
email: peterson@bst.rochester.edu

SUMMARY. Modern high-throughput technologies allow us to simultaneously measure the expressions of a huge number of candidate predictors, some of which are likely to be associated with survival. One difficult task is to search among an enormous number of potential predictors and to correctly identify most of the important ones, without mistakenly identifying too many spurious associations. Mere variable selection is insufficient, however, for the information from the multiple predictors must be intelligently combined and calibrated to form the final composite predictor. Many commonly used procedures over-fit the training data, miss many important predictors, or both. Although it is impossible to simultaneously adjust for a huge number of predictors in an unconstrained way, we propose a method that offers a middle ground where some partial multivariate adjustments can be made in an adaptive fashion, regardless of the number of candidate predictors. We demonstrate the performance of our proposed procedure in a simulation study within the Cox proportional hazards regression framework, and we apply our new method to a publicly available data set to construct a novel prognostic gene signature for breast cancer survival.

KEY WORDS: Micro-array data; prognostic signature; gene selection; Cox proportional hazards regression; censored survival data.

1 Introduction

Modern micro-array technologies now allow us to simultaneously measure the expression of a huge number of different genes and proteins, some of which are likely to be associated with cancer prognosis. While such gene expressions are unlikely to ever replace important clinical covariates, evidence is already mounting that they can provide a significant amount of additional predictive information [1] [3] [6] [7] [8] [9] [10] [13] [12] [14]. The difficult task, then, is to search among an enormous number of potential predictors and to correctly identify most of the important ones, without mistakenly identifying too many spurious associations.

Many commonly used procedures unfortunately over-fit the training data, leading to subsets of selected genes many of which are unrelated to the outcome in the target population, despite appearing predictive in the particular sample of data used for subset selection. Part of the reason for this lies in the extreme biases induced by applying standard estimation methods as if genes had not been picked based on their estimates being larger than those of other genes. Thus, our proposed method not only tackles the issue of identifying important genes but also that of better estimating their relationships with the outcome. Indeed, such improved estimation is a key component to selecting appropriate genes in the first place, since selection is typically based upon estimates. Both gene selection and parameter estimation are critical components in constructing valid prognostic gene signatures, and these tasks should ideally be intertwined.

Some genes might only be useful predictors when used in concert with certain other related genes and/or with clinical covariates, yet the vast majority of available methods are inherently univariate in nature, based only on the marginal associations between each predictor and the outcome. While it is impossible to simultaneously adjust for a huge number of predictors in an unconstrained way, we aim to provide a middle ground where some partial adjustments can be made, even in the face of a huge number of candidate predictors.

One multivariate approach that has been considered by some is to use principal components

analysis (PCA) to construct $n - 1$ or fewer orthogonal *eigengenes*, each of which is a linear combination of all p (typically log-transformed) gene expressions [4]. Associations between these eigengenes and the clinical outcome can then be explored in hopes of constructing prognostic signatures [14]. However, unless some (typically *ad hoc*) approach is implemented to exclude the contributions of the majority of the genes from any given eigengene, no gene selection is performed and the predictive signal in the truly important genes can easily be swamped by the noise of the many others. On the other hand, if many elements of the PCA loadings are thresholded to zero after PCA, the key properties of orthogonality and maximal variation among the first few eigengenes are destroyed, begging the question of why PCA was applied in the first place.

A somewhat related strategy uses the idea of partial least squares (PLS, [15] [2] [4]) to derive predictors as linear combinations of the (log) gene expressions. In its pure form, each PLS predictor gives weight to every gene, like PCA. However, the key difference is that these weights are a function not only of the gene expression matrix but also of the outcome. In fact, the first PLS direction is simply formed by performing univariate regressions of the outcome on each of the p genes and using the resulting unadjusted regression coefficients to form the first linear combination. Thus, genes with stronger marginal associations with the outcome are seemingly appropriately given more weight than are genes that are marginally unrelated to the response. The result is that models based on just a handful of PLS components fit the training data far better than models based on a large number of PCA eigengenes. In fact, PLS fits the training data far *too* well. Indeed, when p is much larger than n , even the very *first* PLS predictor can over-fit the training data in the sense that (1) the magnitude of the estimated coefficients are far too large, on average, (2) the likelihood is increased by more than if all of the truly important genes were included by an oracle, and (3) the predictive performance with independent test data is miserable. Moreover, adding additional PLS components results in an infinite likelihood in short order, saturating the model long before $n - 1$ (or even \sqrt{n}) components can be entered. The number of PLS components is not even close to equivalent to the classical notion of degrees of freedom, given that each component is formed

by estimating p parameters. Variations on PLS, where many of the gene loadings are shrunken to zero via some *ad hoc* gene selection procedure might perform somewhat better [6] [7] [5]. However, in the extreme case where all but one of the gene contributions is shrunken to zero for each PLS predictor, the method becomes very similar to classical forward stepwise selection. And since the stepwise selection approach is already too greedy an algorithm, in the sense that it can be shown to badly over-fit the training data, the utility of these variations on PLS are also in serious doubt.

2 Stagewise Forward Search with Shrinkage Estimation

Our approach is a stagewise forward search algorithm that uses constrained estimation, or “shrinkage,” at every stage. The first step of this algorithm selects precisely the same 1st gene as would both stepwise forward selection or most any univariate approach, i.e. that gene whose contribution to the unconstrained likelihood for the univariate regression is largest. However, the parameter estimate for this gene is then iteratively shrunken toward the null value (which is typically 0) and then temporarily *fixed* at this value at the next stage where the set of candidate genes is again searched for the best gene to add to the model. At each stage, only a single parameter is allowed to freely vary in the likelihood, which helps to alleviate the problems experienced by standard stepwise selection procedures as the number of selected variables grows. Thus, this method has much in common with stagewise forward search algorithms [4] that have been used in machine learning to approximate Tibshirani’s L_1 -penalized least squares *lasso* [11] procedure. However, seemingly subtle differences in such algorithms can result in substantial differences in performance. If the lasso stopping rule is weak, then a large number of unrelated genes are typically selected, while a strict stopping rule results in severe underestimation of the associations between each gene and the outcome (i.e. over-shrinkage or under-fitting), in addition to missing several important genes. In either case, the estimated prognostic gene signature resulting from the lasso-type approach will not perform particularly well.

Our proposed method, which we call the *shrinkstage* method, overcomes some of the difficul-

ties associated with the lasso approach due to our novel, easily implemented estimation method used at each stage. First, rather than constraining the estimate at each stage to be below some small increment threshold as with the lasso, we simply shrink the maximum likelihood estimate at each stage by a given shrinkage factor $0 < \gamma \leq 1$. Next, once each new variable has entered the model, we update all of the parameter estimates in the model, but in a constrained fashion. First, we iteratively further shrink any of the estimates if the reduction in the likelihood compared with unconstrained estimation (of the *single* parameter) falls below a given threshold, implied by an approximate p -value-to-shrink. Once the iterative shrinking stage is complete, we allow for an iterative growth stage, governed by an approximate p -value-to-grow. In particular, we allow each parameter estimate to grow by γ times the difference between the current estimate and the maximum likelihood estimate, provided the increase in the likelihood is sufficiently large. Once the growth stage is complete, we consider adding the next variable, according to an approximate p -value-to-enter, which implies the stopping rule. The result is a method which more fully adjusts for the effects of strong predictors early on than does the forward stagewise approximate lasso method, resulting in not only superior parameter estimates but also better selection of relevant genes. However, the parameter estimates still never grow all the way to the maximum likelihood estimates, which protects against over-fitting, in contrast to stepwise subset selection methods.

3 Simulation Results

We tested the new method proposed in Section 2 by applying it in the critical simulation setting where we know the true generating model. In this section we describe our simulation model, the details of each of the methods we compared, the criteria we used to assess the performance of each method, and the results of our simulation experiment.

We generated 20 independent simulated sets of censored survival data, each containing $n = 125$ independent subjects and $p = 1000$ correlated gene expression levels, with only 12 of the genes associated with survival. Each algorithm was applied to precisely the same 20 simulated data sets.

The distribution of the (log) gene expression levels was multivariate normal with unit variances and an autoregressive correlation structure with maximal correlation $r = 0.5$ between any two adjacent genes. Thus, every gene was positively correlated with every other gene, yet each gene was only strongly correlated with a handful of other genes since the correlations tail off geometrically. The distribution of the i^{th} patient's survival time T_i was taken to be exponential with rate $X_i^T \beta$, where X_i denotes the p -vector of (log) gene expression levels for subject i and β denotes the p -vector of log-hazards, or model parameters. The distribution of the independent random censoring time C was standard exponential (rate 1), which induced approximately 50% censoring in each data set since the observed survival data for each subject is the follow-up time $Y_i = \min(T_i, C_i)$ and the censoring indicator $\Delta_i = I(T_i \leq C_i)$. For simplicity, each element of the parameter vector β contained 1 of 3 possible values: 0 (unrelated gene), 1 (up-regulation increases risk by 2.718-fold per SD), or -1 (down-regulation increases risk by 2.718-fold per SD). The magnitude of 1 was chosen so as to be large enough to be detectable by good procedures yet not so huge as to be easily detectable by even poor procedures. Rather than vary the magnitudes of the coefficients in this particular experiment, we chose to vary only the correlations between the relevant genes, as follows. Out of the 1000 autocorrelated genes, 988 were unrelated to survival, while the following 12 were relevant predictors: 150, 151, 300, 302, 450, 453, 600, 604, 750, 755, 900, 906; and the corresponding 12 nonzero coefficients were 1, 1, -1, -1, 1, 1, -1, -1, 1, 1, -1, -1; thus, risk was associated with up-regulation for 6 genes and down-regulation for another 6. The correlation between (log) gene expressions 150 and 151 was 0.5, that between 300 and 302 was 0.25, and so on, so that the correlation for genes 900 and 906 was only 0.015625. Also note that, e.g., irrelevant gene 301 had correlation 0.5 with both genes 300 and 302, noise gene 602 has correlation 0.25 with both genes 600 and 604, etc.

The following performance criteria were computed for each method applied to each of the 20 simulated data sets. We counted the number of relevant genes identified (out of 12), the number of irrelevant genes additionally selected (out of 988), and the total number of genes selected (or the

size of each model). Since the goal here is not just to identify genes but to construct a complete prognostic gene profile, i.e. a linear combination of (log) gene expression levels that predicts survival, good estimation of the log-hazard-ratios for each gene is also critical. Thus, as a criterion to judge the predictive performance of each method, we used a version of the Mean Squared Prediction Error (MSPE) for the linear predictor of the Cox Proportional hazards model, given by the multivariate distance

$$\text{MSPE} = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta) \tag{1}$$

between the true vector β and the estimated p -vector $\hat{\beta}$, where the latter contains zeroes for every gene that was not selected, and where Σ denotes the covariance matrix of the (log) gene expressions. The more common definition of MSPE would be to replace Σ in equation (1) by its empirical random counterpart $X^T X$, where X denotes the (log) gene expression matrix from an independent *test* sample. While we certainly could have done this, we note that by replacing $X^T X$ by its expectation Σ we reduce the simulation-to-simulation variability of MSPE, while retaining all of its attractive properties, including the fact that it depends on the observed gene expression values in the *training* data only through the parameter estimates $\hat{\beta}$. The MSPE does ignore estimation of the nonparametric baseline hazard function, which is treated as an infinite-dimensional nuisance parameter when fitting the Cox model using the partial likelihood. But the MSPE takes proper account of the covariance Σ of the (log) gene expression values and penalizes for selecting inappropriate genes, missing relevant genes, and for estimation error (overestimating or underestimating each gene's effect). If every parameter were perfectly estimated, MSPE would be 0, so methods with the smallest MSPE provide superior predictions to those with larger MSPE. We also calculated the Root Mean Squared Prediction Error (RMSPE), which is simply the square root of MSPE, since it has the nice interpretation of being the magnitude of a typical prediction error (on the original log-hazard rate scale).

For the shrinkstage method, the following *ad hoc* control parameters were used: the shrinkage

Table 1: Median Performance Over 20 Simulated Sata Sets

Method	MSPE	RMSPE	#true	#false	size
oracle	0.6	0.8	12.0	0.0	12.0
shrinkstage	5.5	2.4	8.5	5.5	14.0
univariate	9.3	3.1	6.0	7.0	13.0
stepwise	13.0	3.6	6.5	8.0	14.5
empty	14.0	3.7	0.0	0.0	0.0

factor γ was 0.85, p -to-enter was 0.015, p -to-shrink was 0.8, and p -to-grow was 0.5. The stepwise selection algorithm was started from the empty model and expanded at each iteration with a p -to-enter of 0.0025; after each variable was entered, the stepwise algorithm was also programmed to drop any previously entered variable with a p -to-drop of 0.0025 so that only those variables whose Likelihood Ratio Test(LRT)-based p -values of 0.0025 or smaller were retained. These nominal p -values for the stepwise algorithm were chosen so as to make the typical size of the resulting model similar to the other procedures. However, it was clear that these values were still actually a bit too high to prevent the stepwise procedure from occasionally selecting *far* too many genes (since after the 16th selected gene, only garbage would typically enter). Thus, an additional stopping rule was imposed to limit the stepwise procedure to selecting no more than n_u^7 (or about 18, on average) genes, where n_u denotes the number of uncensored observations.

In addition the shrinkstage and stepwise methods, the following methods were compared. To provide a guideline as to the worst tolerable MSPE, MSPE was calculated for the *empty* model (selecting no genes); any method doing worse than this (MSPE = 14), e.g. when too many unrelated genes are selected, is obviously a seriously flawed procedure. To provide an envelope as to nearly the best (perhaps, unattainable) performance one might expect, the *oracle* model was fit, including the 12 relevant genes and no others. In order to assess the performance of typical univariate procedures, the following univariate procedure was compared: each parameter was estimated via univariate Cox regression, and only those genes with nominal LRT p -values less than 0.005 were selected; all of their parameters were then simultaneously estimated by including just those selected genes in a single multivariate Cox model. Again, the nominal alpha level of 0.005 was chosen so as

to make the average size of the selected subset close to 12, like the other methods; a lower threshold would choose fewer noise variables but also identify fewer correct genes. The medians over the 20 simulated data sets for all of the methods appear in Table 1.

As is clear from Table 1, our proposed methods perform much better than both stepwise selection and the univariate selection procedures, choosing more of the true predictors (8-11 *vs.* 6-7, typically), selecting fewer of the unrelated genes (1-6 *vs.* 7-8), and attaining lower prediction errors. The typical prediction error (median RMSPE) for stepwise selection (3.6) is 50% larger than that for our shrinkstage method (2.4), e.g. What was somewhat surprising was that even the univariate procedure outperformed the stepwise algorithm here, which probably explains why few researchers are employing such stepwise selection procedures with high-dimensional microarray data. The shrinkstage method can also be improved, e.g. via judicious choices of the tuning parameters, though we note that, in contrast to the stepwise method, its performance was very stable over all 20 simulated data sets, which is encouraging. Altering the tuning parameters of the stepwise and univariate methods to choose smaller models could reduce the prediction error somewhat and select fewer incorrect genes, but there is no way to get either procedure to select as many correct genes as our method and thus no way to reduce the prediction error to the levels we have already achieved.

Although not shown in Table 1, we also began investigating the performance of sequentially adding PCA-derived eigengenes, as well as PLS-type predictors. We observed that even if a large number of PCA-derived eigengenes are included, the models essentially underfit severely, to the point where the MSPE is hardly lower than the empty model. The reason for this seems to be that since each eigengene is a linear combination of 988 noise genes and 12 true predictors, the maximum likelihood coefficient estimates have to be very close to 0 so as to keep the contributions from noise genes suppressed. And while PCA seems to essentially never overfit, PLS often overfits in step one. Adding even a single PLS-derived predictor resulted in grossly overestimated coefficients, with correspondingly poor MSPE that was typically worse than the null model. Thus, we cannot

recommend either of these procedures for use in the context of high-dimensional predictors, at least when the true predictors are relatively low-dimensional, as in our simulation context here.

4 Analysis of Breast Cancer Survival

We applied our novel method to van de Vijver et al’s [12] publicly available breast cancer survival data on $n=295$ subjects with expressions measured on 70 genes that van’t Veer et al [13] previously screened from 24,885 candidate genes. Since data on the vast majority of the 24,885 candidate genes was unavailable to us, we were unfortunately unable to shed any light on which genes beyond the 70 provided might be related to survival. However, even when only considering 70 genes, we observed interesting differences in the results of applying various methods to select the predictive genes and estimate their impact on mortality.

Table 2 displays the relative hazard of mortality per unit change in the standard deviation of the expression of each gene. Only those 32 of the 70 genes that were selected by at least one procedure are shown, whereas the estimated relative hazards are all 1.00 for the other 38 genes. The genes are sorted by their univariate approximate large-sample p -values based on likelihood ratio tests, and the top 16 genes had univariate p -values sufficiently small to satisfy even a 0.05 level test with a Bonferroni correction for 24,885 tests, while 63 of the 70 genes would meet the unadjusted nominal 0.05 level test. The `uni.Bonf` column contains the corresponding *unadjusted* relative hazards for each of the top 16 univariately selected genes. Although Bonferroni-corrected tests are often criticized for being overly “conservative,” we remark that even this might not necessarily be true, in general. The Bonferroni correction is indeed conservative when applied to *exact* p -values based on nonparametric tests, or when each and every p -value is exactly uniformly distributed under the null hypothesis, but these conditions are unfortunately seldom met with gene expression data in practice. In the present context, e.g., one would need to trust the accuracy of the large sample approximation based on only 79 events among 295 subjects out to 6 or more decimal places of the approximate p -value in order to begin to have confidence that the Bonferroni correction is

Table 2: Relative hazards per SD of expression for several methods applied to van de Vijver et al's breast cancer data on n=295 subjects and 70 genes (screened from 24,885), sorted by univariate p -value. 38 of the 70 genes selected by *none* of the methods (all relative hazards = 1.00) not shown.

gene	uni.Bonf	ss	ss.adj	lasso	lasso.adj	step	step.adj	step.back
NM.003981LRa	2.10	1.28	1.00	1.05	1.00	1.00	1.00	1.00
NM.016359LRa	2.13	1.76	2.13	1.37	1.40	2.40	2.25	1.00
Contig38288.RCLRa	1.87	1.25	1.00	1.16	1.00	1.00	1.00	1.00
NM.001809LRa	1.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Contig55725.RCLRa	1.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.014321LRa	1.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.020974LRa	0.51	1.00	1.00	0.95	1.00	1.00	1.00	1.00
NM.004702LRa	1.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.014791LRa	1.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AL137718LRa	1.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Contig48328.RCLRa	0.51	1.00	1.00	0.87	1.00	1.00	1.00	1.00
NM.016448LRa	1.83	1.00	1.00	1.04	1.08	1.00	1.00	1.95
Contig28552.RCLRa	1.76	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Contig46218.RCLRa	1.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.005915LRa	1.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.000849LRa	0.57	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NM.002916LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.54
Contig46223.RCLRa	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.71
NM.020188LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.03
AF052162LRa	1.00	1.46	1.41	1.16	1.12	1.55	1.47	1.61
X05610LRa	1.00	1.42	1.34	1.19	1.08	1.60	1.44	1.72
Contig63649.RCLRa	1.00	1.34	1.33	1.19	1.13	1.47	1.42	1.39
NM.003875LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.63
AF201951LRa	1.00	1.00	1.00	0.90	0.95	1.00	1.00	1.00
NM.003862LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.37
AF257175LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.59
NM.006117LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.78
NM.002073LRa	1.00	1.00	1.00	1.00	1.02	1.00	1.00	1.00
AF055033LRa	1.00	1.29	1.38	1.09	1.21	1.37	1.00	1.67
NM.000599LRa	1.00	1.00	1.00	1.00	1.00	1.00	1.42	1.00
Contig32125.RCLRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.36
Contig32185.RCLRa	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70

actually conservative here, even setting aside the very serious issue of confounding that plagues all univariate methods.

With the exception of `uni.Bonf`, the rest of the estimated relative hazards in Table 2 are all *adjusted* in some fashion for the other selected genes with non-unity relative hazards in the same column, similar to any multivariate Cox proportional hazards regression model. Before considering any genes, we first selected a clinical-only Cox model based only on the 4 available clinical indicators: ESR1 (1 = estrogen receptor alpha expression, 0 = no), St.Gallen criterion (1 = low risk, 0 = high risk), NIH criterion (1 = low risk, 0 = high risk), and lymph node status from the pathology report (1 = positive, 0 = negative). Both ESR1 (HR = 0.332, $p < 0.000005$) and the St. Gallen (HR = 0.205, $p = 0.036$) indicators were associated with reduced mortality when included in the same model, despite their heavy interdependence given that ESR1 is necessary for St. Gallen, while neither the NIH criterion nor the lymph node status appeared to predict mortality risk, with or without adjusting for other predictors ($p > 0.15$). The `ss.adj` and `ss` columns contain the results of applying our novel shrinkstage algorithm, with and without forcing the ESR1 and St. Gallen clinical indicators into the model, respectively. The five genes selected by both of these procedures were `NM.016359LRa`, `AF052162LRa`, `X05610LRa`, `Contig63649.RCLRa`, and `AF055033LRa`, and we suggest that these particular 5 genes are thus deserving of further investigation. When the clinical predictors were excluded from the model, `NM.003981LRa` and `Contig38288.RCLRa` were additionally selected by our procedure; our interpretation of this is that the information in these 2 genes is highly correlated with the 2 clinical predictors, and thus these genes add very little prognostic information to the available clinical predictors. Of the 5 genes we selected from the 70 available, adjusted for the clinical predictors, it is interesting but not entirely surprising to us that only 1 fell in the set of the top 16 univariately ranked genes. Looking at the correlations between `NM.016359LRa` and the other top 15 univariately ranked genes, we found that 10 of the 15 pairwise correlations fell between 0.6 and 0.8, indicating a lot of redundant information –not obviously helpful for predicting mortality risk. Finally, we re-assessed the independent contributions of ESR1 and

the St. Gallen indicator via likelihood ratio-type tests to remove each, while leaving the other in the model, now adjusted for our 5 selected genes via an offset to force the gene coefficients to remain at their shrinkstage-estimated levels (resulting in a somewhat lower likelihood from the submodels than would be obtained if we allowed the gene coefficients to be re-estimated in each submodel). While ESR1 was still significant ($p = 0.0077$), the St. Gallen indicator was not ($p = 0.45$) unless tested jointly with ESR1 ($p = 0.0165$ on 2 df), adjusted for the 5 selected genes. This implies that whatever predictive information was contained in the St. Gallen indicator was apparently captured by our low-dimensional gene profile along with ESR1. However, since our development of the gene profile was predicated on adjusting for both ESR1 and the St. Gallen criterion, we did not remove St. Gallen but rather simply note the interesting loss of significance, after genetic adjustment. We also re-assessed the potential contributions of lmyph node status and the NIH criterion, since it is possible that they could become useful in tandem with our selected gene profile, but there was still no evidence that either improved the model in any way ($p > 0.5$).

We programmed a custom stagewise approximation to the lasso procedure in the Cox model framework, with an option to force variables into the model without constraints on their coefficients, and with a default novel stopping rule based on a nominal 0.05 level likelihood ratio test to enter a variable or increase the coefficient on a previously entered variable. We ran this procedure first just on the 70 genes (`lasso`) and again with the 2 clinical indicators additionally forced into the model (`lasso.adj`). Without adjusting for the clinical predictors, the lasso selected 12 genes, 6 of which were among the top 16 univariately selected genes, which was rather different from our shrinkstage selections. In fact, the unadjusted lasso selected several genes that no other multivariate method did, with most of them strongly univariately associated with survival. We believe that this is because at the early stages of the procedure the selections are not sufficiently adjusted for other important multivariate predictors, causing the lasso to start off acting much like a univariate selection procedure. Adjusted for the ESR1 and St. Gallen indicators, however, the lasso selected 8 genes, with only 2 among the top 16 univariate genes, and the overall profile was quite similar

to `ss.adj` but with a few more selected genes and with more attenuated relative hazard estimates. This fits with our anecdotal simulation experience with the lasso in the Cox model (not shown) in that it appears to over-shrink parameter estimates of true predictors unless the stopping rule is very weak, while selecting too many false predictors unless the stopping rule is very strong. But it appears that adjusting for the relevant clinical predictors did improve the lasso's performance.

We also applied 3 versions of stepwise Cox regression, all using a nominal 0.01 level likelihood ratio test to stop: `step` represents a stepwise procedure starting from the null model, and without forcing the clinical predictors in the model; `step.adj` starts from the model with the ESR1 and St. Gallen criteria forced into the model; and `step.back` starts from the full model with all 70 genes, with both clinical predictors forced into the model. Interestingly, the unadjusted stepwise search selected exactly the same 5 genes as `ss.adj`, though 4 of the 5 relative hazard estimates were slightly stronger. Adjusting for the clinical indicators resulted in slightly more attenuated relative hazard estimates for the genes, and NM.000599LRa was selected instead of AF055033LRa, which hardly matters given that the correlation between these 2 gene expressions was 0.975 (indicating that either is probably nearly as good a predictor). Starting from the full 72-predictor model and stepping back resulted in selecting 14 genes, 8 of which were selected by none of the other procedures. Moreover, this was the *only* procedure not to select NM.016359LRa. Overall, we do not believe that `step.back` performed well here, and we cannot recommend it as a general approach; certainly such an approach is never even an option when the number of candidate predictors exceeds the number of uncensored observations. The apparently reasonable performance of the forward stepwise searches is probably in part due to the relatively limited number of available predictors here, as it comes as somewhat of a surprise in comparison with our simulation results that suggest that forward stepwise searches typically miss many important predictors and often select far too many irrelevant genes. In fact, the high concordance of our novel method with the forward stepwise search here leads us to believe that there might be several important predictors among the 24,815 genes that were pre-screened and unavailable to us for this analysis.

Finally, our definition of a *prognostic signature* is not merely a set of genes but rather a complete equation that precisely defines a single continuous risk score, based here on the linear predictor of the Cox model (for the log-relative-hazard). For the `ss.adj` procedure that we advocate, the continuous risk score is simply calculated as:

$$\begin{aligned} \text{risk - score} = & -0.626(\text{ESR1}) - 0.682(\text{St.Gallen}) + 2.908(\text{NM.016359LRa}) + 1.160(\text{AF055033LRa}) \\ & + 2.053(\text{AF052162LRa}) + 1.186(\text{Contig63649.RCLRa}) + 1.844(\text{X05610LRa}) \end{aligned}$$

Note that the coefficients on the gene expressions in the risk score equation above correspond with the original gene expression scales, not standardized by their sample standard deviations, whereas the relative hazards reported in Table 2 are per unit standard deviation of change in order to better compare the relative impact of each gene. The above risk score could potentially be used to risk stratify future subjects based on these 5 genes and 2 clinical indicators, with lower risk scores having superior prognosis –but until independently validated in a future study, caution should be exercised in interpretation. For the 295 subjects used to develop this risk score, the distribution of the risk score appeared approximately normal with mean -0.881, standard deviation 1.264, 75th percentile 0.039, and 95th percentile 1.131. So, roughly speaking, positive risk scores correspond to those in the upper quartile of risk. Van de Vijver et al ultimately classified each subject as having “good” or “poor” prognosis by thresholding a subject-specific correlation coefficient of their 70-gene profile with the average gene profile of “good outcome patients,” defined as all patients with no observed distant metastasis during follow-up, regardless of the length of follow-up. We calculated the sample correlation coefficient between our risk scores and their continuous measure to be -0.783 ($p < 0.000001$), and a scatter plot revealed the relationship to be very linear. Thus, although the agreement in prognosis not perfect and it is unclear which might be superior, the concordance is relatively high, even though our signature is based on 14 times fewer genes. Since the coefficients on all 5 of our selected genes are positive, higher expressions of these genes appear to correspond with increased risk of mortality. Since the coefficients on both clinical predictors are

negative, ESR1 and the St. Gallen indicator indeed correspond with lower risk, as their definitions imply. Having ESR1 corresponds with a 46.5% reduction in risk, all other gene expressions being the same, while meeting the St. Gallen criteria corresponds with a full 73.0% risk reduction, since any woman satisfying the St. Gallen criteria also has ESR1 (and so the coefficients on ESR1 and St.Gallen in the risk score equation are summed, prior to exponentiating to obtain the relative hazard). Thus, any prognostic signature that ignores these available clinical predictors would seem to be quite incomplete and potentially misleading. However, if our risk score equation reasonably approximates the true log-relative hazard, it does imply that up-regulation of 3 or more of the 5 risk genes by even a single standard deviation could more than offset the protective effects associated with the St. Gallen criterion.

5 Discussion

We introduced a novel method of constructing a prognostic gene signature based on censored survival data, with multivariate adjustments for clinical covariates as well as for multiple selected genes. We chose to reduce the problem to that of model selection and parameter estimation in the Cox proportional hazards model framework in the context of high-dimensional data, where most standard model selection strategies and estimation methods typically break down rather badly. Given this choice of model framework, our procedure inherits both strengths and weaknesses typically associated with the Cox model. For example, if interactions are to be modeled, they must first be explicitly defined by the user and included as additional candidate predictors, and doing this can quickly expand the number of candidate predictors to enormous numbers, making the model selection problem that much harder. In this paper, we chose to assume no interactions, including interactions with time, though such enrichments could be used to partially relax the proportional hazards assumption.

We compared our procedure to stepwise selection and to fitting a Cox model to a univariately screened set of predictors. In a simulation setting where the Cox model is correct, provided all of the

true predictors are included in the model, we demonstrated the clear superiority of our procedure to these simple standard approaches, both of which were observed to perform quite poorly. However, in future work we plan to investigate a wider set of simulation models, as well as to compare some of the more recently introduced competing approaches, and we look forward to further improving our method, e.g. given its current dependence on *ad hoc* tuning parameters.

We applied our novel method to breast cancer survival data, using a data set with a relatively large number of subjects, and we identified a much simpler prognostic signature compared with previous attempts by others. Our profile focuses on just 5 key genes, and it includes already established clinical predictors, which we argue is the most sensible way to proceed. Our results suggest that profiles based on 70 or more genes might be overly complicated, while only a handful might be needed for effective prognosis. Unfortunately we did not have access to the expression data on the vast majority of the nearly 25,000 genes, and so we are left to wonder how many of them might also be associated with survival, adjusted for clinical predictors and other genes.

6 Acknowledgements

This work was partly supported by Concept Award W81XWH-04-1-0714 from the Breast Cancer Research Program of the Congressionally Directed Medical Research Programs run by the US Department of Defense via the US Army Medical Research and Materiel Command.

7 Literature Cited

References

- [1] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, and Yakhini Z (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology* **7**:559–584.
- [2] Garthwaite PH (1994). An interpretation of partial least squares. *Journal of the American Statistical Association* **89**:122–127.

- [3] Golub, TR, Slonim, DK, Tamayo P, *et al* (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**:531–537.
- [4] Hastie T, Tibshirani R, and Friedman J (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
- [5] Li H and Gui J (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *ISMB04/Bioinformatics* (in press).
- [6] Nguyen DV and Roche DM (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**:39–50.
- [7] Park PJ, Tian L, and Kohane IS (2002). Linking expression data with patient survival times using partial least squares. *Bioinformatics* **18**:1625–1632.
- [8] Pomeroy SL, Tamayo P, Gaasenbeek M, *et al* (2001). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**:24.
- [9] Rosenwald A, Wright G, Wiestner A, *et al* (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**:185–197.
- [10] Sorlie T, Perou CM, Tibshirani R, *et al* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS* **98**:10869–10874.
- [11] Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**:267–288.
- [12] van de Vijver MJ, He YD, van't Veer LJ, *et al* (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347**:1999–2009.
- [13] van't Veer LJ, Dai H, van de Vijver MJ, *et al* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**:530–536.

- [14] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, *et al* (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS* **98**:11462–11467.
- [15] Wold H (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah PR (Ed.), pp. 391–420. Academic Press.