Partial least squares Cox regression on genomic data handling additional covariates

Ståle Nygård
Department of Mathematics, University of Oslo
P.O. Box 1053 Blindern, N-0316 Oslo, Norway
e-mail: staaln@math.uio.no

Ørnulf Borgan
Department of Mathematics, University of Oslo
P.O. Box 1053 Blindern, N-0316 Oslo, Norway
e-mail: borgan@math.uio.no

Ole Christian Lingjærde
Department of Informatics, University of Oslo
P.O. Box 1080 Blindern, N-0316 Oslo, Norway
e-mail: ole@ifi.uio.no

Hege Leite Størvold
Department of Informatics, University of Oslo
P.O. Box 1080 Blindern, N-0316 Oslo, Norway
e-mail: hegelst@ifi.uio.no

Abstract
There exist many methods for survival prediction from high-dimensional
genomic data. Most of them combine the Cox proportional hazards model with some dimension reduction estimation technique, like partial least squares (PLS). For PLS it is not obvious how it should be applied to the Cox model, and different approaches have been suggested. The perhaps most reasonable one, Park et al. (2002), uses a reformulation of the Cox likelihood to a Poisson type likelihood, thereby enabling estimation by iteratively reweighted partial least squares for generalized linear models. We present a modified version of the method of Park et al. (2002), which estimates the baseline hazard and the gene effects in separate steps. Our approach has the advantages of leading to PLS directions that have more reasonable biological interpretations, providing estimates of survival probabilities for new patients and enabling a much and less memory-demanding estimation procedure. In addition our method allows for incorporation of lower-dimensional non-genomic variables like disease grade and tumor thickness. Applying our method to two different microarray gene expression breast cancer data sets, one with additional non-genomic covariates, shows that our method gives at least as good predictions as the method of Park et al. (2002), and that there is a lot to be gained by including gene expressions together with the clinical variables.

Key words: Cox regression; Dimension reduction; Gene expression data; High-dimensional data; Partial least squares; Survival prediction

1 Introduction

Predicting the outcome of a disease based on genome-wide data is an important application of microarrays and other high-throughput sources. One particular instance of this problem seeks to link genome-wide molecular profiles of tumors to the time of some disease specific event such as death or recurrence of disease; frequently referred to by the technical term failure time or survival time. A classical statistical model for survival data that has been explored recently in this context by multiple authors is the Cox proportional hazards model.

In conjunction with the genomic data, information on demographic and clinical variables are often available. Assume that we have two covariate vectors $\mathbf{x} = (x_1, \ldots, x_p)^T$ and $\mathbf{z} = (z_1, \ldots, z_q)^T$, the first corresponding to genome-wide data such as microarray gene expression measurements, and the second corresponding to additional demographic and clinical variables (e.g. such as sex, age, and tumor thickness) that are few in number and not
reasonable to treat on the same footing as the genomic data. In a Cox model that links the covariate vectors $\mathbf{x}$ and $\mathbf{z}$ with the patients' failure times, one assumes that the hazard rate, i.e. the instantaneous risk of failure at time $y$ takes the form

$$\lambda(y \mid \mathbf{x}, \mathbf{z}) = \lambda_0(y)e^{\mathbf{x}^T \beta + \mathbf{z}^T \gamma}.$$  \hspace{1cm} (1)

Here $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\gamma = (\gamma_1, \ldots, \gamma_q)^T$ are vectors of regression coefficients describing the effects of the gene expressions and the non-genomic covariates, respectively, and $\lambda_0(y)$ is a baseline hazard common to all patients. The baseline hazard is assumed to be an arbitrary non-negative function, and it corresponds to the hazard rate of an individual with all $x_j$ and $z_j$ equal to zero.

Usually, the number of gene expressions $p$ in (1) will by far exceed the number $n$ of patients in the study ($p \gg n$). Some form of dimension reduction, or regularization, is then required to obtain useful parameter estimates. For the classical case of ordinary linear regression, many methods exist for this purpose; including methods that select a subset of the covariates to be used in the regression, principal components regression (PCR), partial least squares (PLS), ridge regression and the lasso; see e.g. Hastie et al. (2001) for a review. Many of these methods have been applied to Cox’s regression model (1) for the case where there are no additional covariates (i.e. where $\mathbf{z} = \mathbf{0}$). Bair and Tibshirani (2004) and Bair et al. (2006) proposed "supervised PCR", a method that uses PCR on a subset of the genes for which the Cox scores exceed a certain threshold, while van Houwelingen et al. (2006) showed how ridge regression can be adopted to Cox’s model using a penalized likelihood approach. Nguyen and Rocke (2002) applied PLS to Cox regression by substituting the original vector of gene expressions in (1) by a lower-dimensional vector obtained through bivariate PLS regression of the censored failure times and the censoring indicators on $\mathbf{x}$.

Following Whitehead (1980), Park et al. (2002) expressed the likelihood for the Cox model as a Poisson type likelihood, thereby enabling application of the iterative reweighted partial least squares (IRPLS) algorithm (Marx, 1996), in which the weighted least squares step in the iterative reweighted least squares (IRLS) algorithm (McCullagh and Nelder 1989) is replaced by weighted partial least squares.

Neither of the mentioned dimension reduction methods explicitly state how estimation should be performed in the presence of the additional non-genomic covariates $\mathbf{z}$, but for most of them fairly straightforward solutions exist. For subset selection a natural approach could be to start out by estimating the effect of the non-genomic covariates $\mathbf{z}$, and then additionally
include a subset of the genomic covariates $x_j$ such that the prediction is
optimally improved. For PCR one could derive the principal components
based on (all or a subset of) the genomic covariates, before performing an
ordinary Cox regression on these components together with the non-genomic
covariates. The natural extension for ridge Cox regression would be to pe-
nalize the estimation of $\beta$ in (1), but not that of $\gamma$. For PLS, however, it
is not so obvious how the additional covariates $z$ should be treated, at least
not in the framework of the existing methods. A symmetrical treatment of
the covariates $x$ and $z$, would lead to PLS components being mixtures of
 genomic and non-genomic covariates, which would lack interpretation, and
presumably also result in poor prediction.

In this paper, generalizing the work of Park et al. (2002), we suggest
a version of PLS for Cox’s model which imposes dimension reduction only
on the gene expressions. In Park et al. (2002), even the baseline hazard in-
crements are treated on an equal footing with the genomic variables, while
our method keeps the estimation of the baseline hazard, as well as the non-
genomic variables if available, outside the dimension reduction of the PLS
estimation. This is important, since it allows us to predict survival probabil-
ities for given values of the gene expressions and the additional covariates.
Another advantage of our approach is that we can use the original data ma-
trix as input for estimation instead of an extended matrix which is required
for the method of Park et al. (2002), thus increasing the speed and decreasing
the computer memory demand, which can be crucial for these kinds of
data.

The outline of the paper is as follows. In Section 2 we show how iterative
rewighted least squares can be applied to estimate regression coefficients
for the two sets of covariates and the baseline hazard in the Cox model
(1). Then in Section 3 we show how the least squares step in this IRLS
algorithm can be replaced by PLS, giving an IRPLS algorithm similar to
the one of Marx (1996), but modified so that the baseline hazard as well as
non-genomic variables are held outside the PLS regression on the genomic
covariates. In Section 4 we use two different microarray gene expression
breast cancer data sets, one with additional covariates, to make comparisons
between the performance of (i) our and Park’s method and (ii) our method
with and without gene expression covariates included in the model. Finally
we discuss the method and conclude in Section 5.
2 An IRLS algorithm for Cox regression

Assume that we have censored survival data, \((y_i, \delta_i, x_i, z_i); \ i = 1, \ldots, n\); for \(n\) individuals. Here \(y_i\) is the observed or censored failure time for the \(i\)th individual, \(\delta_i\) is an indicator for whether the failure time is observed \((\delta_i = 1)\) or censored \((\delta_i = 0)\), and \(x_i = (x_{i1}, \ldots, x_{ip})^T\) and \(z_i = (z_{i1}, \ldots, z_{iq})^T\) are two vectors of covariates.

The full likelihood for the Cox model in (1) may be written

\[
L = \prod_{i=1}^{n} \left[ (\lambda_0(y_i)e^{x_i^T\beta + z_i^T\gamma})^{\delta_i} \exp\left\{ -\lambda_0(y_i)e^{x_i^T\beta + z_i^T\gamma} \right\} \right],
\]

where \(\lambda_0(y)\) is the cumulative baseline hazard. For the absolutely continuous case the likelihood can be made arbitrarily large by letting \(\lambda_0(y)\) be zero except from close to the observed failure times, where we let it peak higher and higher. However, by considering an extended model, where the cumulative baseline hazard \(\Lambda_0(t)\) may be any non-negative, non-decreasing function, the full likelihood is maximized if all probability mass is put at the distinct observed failure times \(t_1 < t_2 < \cdots < t_D\) (Johansen 1983). With this extension \(\Lambda_0(y_i) = \sum_{j: t_j \leq y_i} \Delta\lambda_0(t_j)\), where \(\Delta\lambda_0(t_j)\) is the increment of the cumulative baseline hazard at time \(t_j\), and the full likelihood takes the form

\[
L = \prod_{i=1}^{n} \left[ (\Delta\lambda_0(y_i)e^{x_i^T\beta + z_i^T\gamma})^{\delta_i} \exp\left\{ -\lambda_0(y_i)e^{x_i^T\beta + z_i^T\gamma} \right\} \right].
\]

The increments \(\Delta\lambda_0(t_j)\) of the baseline cumulative hazard and the regression coefficient vectors \(\beta\) and \(\gamma\) are estimated by maximizing the likelihood (2).

For given values of \(\beta\) and \(\gamma\) the values of the baseline hazard increments that maximize the likelihood are given by

\[
\Delta\hat{\lambda}_0(t_j | \beta, \gamma) = \frac{d_j}{\sum_{k:y_k \geq t_j} e^{x_k^T\beta + z_k^T\gamma}},
\]

where \(d_j = \sum_{k:y_k = t_j} \delta_k\) is the number of individuals failing at \(t_j\). If (3) is inserted in (2) one gets a profile likelihood for \(\beta\) and \(\gamma\) that is proportional to Breslow's modification of Cox's famous partial likelihood

\[
Pl(\beta, \gamma) = \prod_{j=1}^{D} \prod_{y_k = t_j} \left( \frac{e^{x_k^T\beta + z_k^T\gamma}}{\sum_{i:y_i \geq t_j} e^{x_i^T\beta + z_i^T\gamma}} \right)^{\delta_i}.
\]

Thus maximization of the likelihood (2) may be performed in two steps: first the logarithm of (4) is maximized w.r.t. \(\beta\) and \(\gamma\) using some iterative
maximization algorithm (e.g. Newton-Raphson), and then these maximum likelihood estimates are inserted in (3) to obtain the maximum likelihood estimates for the increments of the baseline hazard. This is the common approach for estimation in Cox’s regression model.

For our purpose, which is to set the scene for the partial least squares method of Section 3, it is however more appropriate to maximize the likelihood simultaneously with regard to the baseline hazard increments \( \hat{\Delta} \lambda_0(t_j) \) and the vectors of regression coefficients \( \beta \) and \( \gamma \). This can be achieved by first setting initial values of the vectors of regression coefficients, e.g. by letting \( \hat{\beta}^{(0)} = 0 \) and \( \hat{\gamma}^{(0)} = 0 \), and then iterating the following steps for \( k = 0, 1, 2, \ldots \) until the the changes in the logarithm of the likelihood (2) are small:

- Get updated estimates \( \Delta \hat{\lambda}_0^{(k)}(t_j) \) of the baseline hazard increments using (3) with parameter vector estimates \( \hat{\beta}^{(k)} \) and \( \hat{\gamma}^{(k)} \).
- Get updated estimates \( \hat{\beta}^{(k+1)} \) and \( \hat{\gamma}^{(k+1)} \) of \( \beta \) and \( \gamma \) based on the current estimates of the baseline hazard increments and the regression parameter vectors.

We use the IRLS algorithm to update the two vectors of regression coefficients. Specifically, given current estimates \( \Delta \hat{\lambda}_0^{(k)}(t_j) \), \( \hat{\beta}^{(k)} \) and \( \hat{\gamma}^{(k)} \) of the baseline hazard increments and the regression parameter vectors, we will get updated estimates of \( \beta \) and \( \gamma \) in two steps: First we use the IRLS algorithm to get an updated estimate of \( \beta \) treating the current estimates \( \Delta \hat{\lambda}_0^{(k)}(t_j) \) and \( \hat{\gamma}^{(k)} \) of the baseline hazard increments and the vector of regression coefficients \( \gamma \) as fixed quantities (offsets) in the regression. Then we update the estimate of \( \gamma \) treating \( \Delta \hat{\lambda}_0^{(k)}(t_j) \) and \( \hat{\beta}^{(k)} \) as offsets. The motivation for doing this in two steps – which for the ordinary IRLS algorithm would be done in one single step – is that we in Section 3 will substitute the IRLS updating step of the high-dimensional coefficient vector \( \beta \) with PLS estimation, but still rely on IRLS to update the low-dimensional coefficient vector \( \gamma \).

To describe how the IRLS updating steps for the vectors of regression coefficients \( \beta \) and \( \gamma \) are performed, we need some notation. First we write

\[
\hat{\beta}^{(k)}_i = \hat{\lambda}_0^{(k)}(y_i) e^{x_i^T \hat{\beta}^{(k)} + x_i^T \hat{\gamma}^{(k)}}
\]

for the cumulative hazard for individual \( i \) (evaluated at its censored survival time \( y_i \)) under the current estimates of the baseline hazard increments and the regression parameter vectors. Then we introduce the "pseudo observa-
tions".
\[
\hat{f}_i^{(k)} = x_i^T \hat{\beta}^{(k)} + \frac{1}{\hat{\mu}_i^{(k)}} (\delta_i - \hat{\mu}_i^{(k)}),
\]
and
\[
\hat{g}_i^{(k)} = z_i^T \hat{\gamma}^{(k)} + \frac{1}{\hat{\mu}_i^{(k)}} (\delta_i - \hat{\mu}_i^{(k)}).
\]
In the appendix, we show that the IRLS updating of \( \beta \) is performed by regressing the "pseudo observations" \( \hat{f}_i^{(k)} \) on the covariates \( x_i \) using weighted least squares with weights \( \hat{\mu}_i^{(k)} \). Similarly the estimate of \( \gamma \) is updated by weighted linear regression of the "pseudo observations" \( \hat{g}_i^{(k)} \) on the covariates \( z_i \).

In order to describe the IRLS updating steps in more details, it is convenient to use vector and matrix notation. To this end we introduce the vector \( \delta = (\delta_1, \ldots, \delta_n)^T \) of failure indicators for the individuals, the vector \( \hat{\mu}^{(k)} = (\hat{\mu}_1^{(k)}, \ldots, \hat{\mu}_n^{(k)})^T \), and the diagonal matrix \( \hat{\gamma}^{(k)} = \text{diag}(\hat{\mu}^{(k)}) \) that has the estimated cumulative hazards (5) on the diagonal. Then the vector of "pseudo observations" (6) takes the form
\[
\hat{f}^{(k)} = X \hat{\beta}^{(k)} + (\hat{\gamma}^{(k)})^{-1} (\delta - \hat{\mu})^{(k)},
\]
while the vector of "pseudo observations" (7) may be written
\[
\hat{g}^{(k)} = Z \hat{\gamma}^{(k)} + (\hat{\gamma}^{(k)})^{-1} (\delta - \hat{\mu})^{(k)}.
\]
Here \( X \) is the \( n \times p \) matrix with rows \( x_i^T \), while \( Z \) is the \( n \times q \) matrix with rows \( z_i^T \). The IRLS updating step for the vector of regression coefficients \( \beta \) may then be given by (cf. the appendix):
\[
\hat{\beta}^{(k+1)} = (X^T \hat{\gamma}^{(k)} X)^{-1} X^T \hat{\gamma}^{(k)} \hat{f}^{(k)}.
\]
In a similar manner the IRLS updating step for \( \gamma \) is given by
\[
\hat{\gamma}^{(k+1)} = (Z^T \hat{\gamma}^{(k)} Z)^{-1} Z^T \hat{\gamma}^{(k+1)} \hat{g}^{(k)}.
\]

The complete modified IRLS algorithm for simultaneous estimation of the vector of regression coefficients and the baseline hazard increments is summarized in Table 1. For simplicity, we have omitted the index \( k \) for the iteration number in this summary.
Table 1: Summary of the modified IRLS method for simultaneous estimation of regression coefficients and baseline hazard in Cox’s regression model.

1) Set initial values of the parameter vectors by letting \( \hat{\beta} = 0 \) and \( \hat{\gamma} = 0 \).
2) Iterate until the changes in the logarithm of the likelihood (2) are small:

(a) Set \( \Delta \hat{\lambda}_0(t_j) = d_j / \sum_{i: y_i \geq t_j} e^{x_i^T \hat{\beta} + x_i^T \hat{\gamma}} \).
(b) Let \( \hat{\mu} \) be the vector with elements \( \hat{\mu}_i = \sum_{j: t_j \leq y_i} \Delta \hat{\lambda}_0(t_j) e^{x_i^T \beta + x_i^T \gamma} \).
(c) Set \( \hat{V} = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_n) \).
(d) Set \( \tilde{f} = X\hat{\beta} + \hat{V}^{-1}(\delta - \hat{\mu}) \), and update the estimate of \( \hat{\beta} \) using weighted least squares regression: \( \hat{\beta} = (X^T \hat{V} X)^{-1} X^T \hat{V} \tilde{f} \).
(e) Set \( \tilde{g} = Z\hat{\gamma} + \hat{V}^{-1}(\delta - \hat{\mu}) \), and update the estimate of \( \hat{\gamma} \) using weighted least squares regression: \( \hat{\gamma} = (Z^T \hat{V} Z)^{-1} Z^T \hat{V} \tilde{g} \).

3 PLS Cox regression

It is a requirement for the modified IRLS algorithm of the previous section that the number of failure times exceed the number of covariates. For genomic data this is seldom the case, however, and some dimension reduction technique is called for. We will follow the approach of Marx (1996) and modify the IRLS algorithm of Table 1 by replacing the weighted least squares regression estimation of \( \eta \) in step (d) with a weighted partial least squares (PLS) regression.

Note that the algorithm of Table 1 updates the regression coefficient vectors \( \beta \) and \( \gamma \) and the baseline hazard in three separate steps, which allows us to apply PLS only to estimate \( \beta \), or the vector of linear predictors \( (\eta_1, \ldots, \eta_n)^T = \eta = X\beta \), which is the quantity used in the Marx’s algorithm. This approach is useful since high-dimensional genomic data often exist in conjunction with non-genomic data of lower dimension, and we find it reasonable to leave the estimation of the latter outside the dimension reduction provided by PLS. Further \( \eta \) and \( \gamma \) ought to be estimated separately from the baseline hazard, which is yet another distinct quantity.

Our iteratively reweighted partial least squares (IRPLS) algorithm for Cox regression is summarized in Table 2. Here (d1)-(d5) are the PLS steps that replace step (d) in the algorithm of Table 1. Steps (d2)-(d4) are weighted versions of the steps in the classical PLS algorithm (e.g. Martens and Næs 1989, Frame 3.4), except that we only do a partial update of \( \hat{\eta} \) in step (d4) to avoid convergence problems. To determine \( c_m \) in step (d4) we perform a line
Table 2: Summary of the iteratively reweighted partial least squares algorithm for Cox’s regression model.

1. Initialize by setting \( \hat{\eta} = 0 \) and \( \hat{\gamma} = 0 \).
2. For \( m = 1, \ldots, s \), where \( m \) is the number of PLS-components iterate until the changes in \( \hat{\mu} \) are small:
   (a) Set \( \Delta \hat{\lambda}_0(t_j) = d_j / \sum_{k:\tilde{y}_k \geq t_j} e^{\tilde{y}_k + z^T \hat{\gamma}} \).
   (b) Let \( \hat{\mu} \) be the vector with elements \( \hat{\mu}_i = \sum_{j: y_j \leq \tilde{y}_i} \Delta \hat{\lambda}_0(t_j) e^{\tilde{y}_i + z_i^T \hat{\gamma}} \).
   (c) Set \( \hat{V} = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_n) \).
   (d1) Set \( \hat{f}_0 = \hat{\eta} + \hat{V}^{-1}(\delta - \hat{\mu}) \).
   (d2) Set \( \hat{E}_0 = \mathbf{X} - \hat{x}^T \), where \( \hat{x} = (\mathbf{X}^T \hat{V} 1)/(1^T \hat{V} 1) \) is the vector of weighted averages of the columns of \( \mathbf{X} \) using weight matrix \( \hat{V} \).
   (d3) For \( k = 1, 2, \ldots, m \), set:
      i. \( \hat{w}_k = \hat{E}_{k-1}^T \hat{V} \hat{f}_{k-1} \)
      ii. \( \hat{\tilde{t}}_k = \hat{E}_{k-1} \hat{w}_k \)
      iii. \( \hat{\gamma}_k = (\hat{f}_{k-1}^T \hat{V} \hat{\tilde{t}}_k)/(\hat{t}_k^T \hat{V} \hat{\tilde{t}}_k) \)
      iv. \( \hat{f}_k = \hat{f}_{k-1} - \hat{\tilde{t}}_k \hat{\gamma}_k \)
      v. \( \hat{p}_k = (\hat{E}_{k-1}^T \hat{V} \hat{\tilde{t}}_k)/(\hat{t}_k^T \hat{V} \hat{\tilde{t}}_k) \)
      vi. \( \hat{E}_k = \hat{E}_{k-1} - \hat{\tilde{t}}_k \hat{p}_k \)
   (d4) Update \( \hat{\eta} \) by \( \hat{\eta} = c_m \hat{T}^{(m)} \hat{\eta} + (1 - c_m) \hat{\eta} \), where \( \hat{T}^{(m)} = (\hat{t}_1, \ldots, \hat{t}_m) \) and \( \hat{\eta}^{(m)} = (\hat{\eta}_1, \ldots, \hat{\eta}_m)^T \), and \( c_m \in (0, 1] \) is a weight maximizing the likelihood, found by performing a line search (see text for details).
   (d5) Set \( \tilde{\eta} = \tilde{\eta} - \frac{1}{n} \sum_{i=1}^n \hat{\eta}_i \).
   (e) Set \( \hat{g} = \mathbf{Z} \hat{\gamma} + \hat{V}^{-1}(\delta - \hat{\mu}) \), and update the estimate of the vector of regression coefficients \( \gamma \) using the weighted least squares regression (11).
search and use the value of \( c_m \in (0, 1) \) that maximizes the total likelihood (2) (with \( x_i^T \beta \) replaced by \( \eta_i \)). Specifically we do this by making a uniform grid of 10 values between 0 and 1 and finding the value of \( c \) that maximizes the total likelihood. We then make a finer tuning by repeating this for a new uniform grid with 50 values locally around the \( c \) found in the first line search. Note that on convergence \( \hat{\eta} = c_m \hat{F}(m) \hat{q}^{(m)} + (1 - c_m) \hat{\eta} \), implying that we then have \( \hat{\eta} = \hat{F}(m) \hat{q}^{(m)} \), corresponding to a full update.

Some further comments on the algorithm of Table 2, and a comparison with the PLS-algorithm of Marx (1996, page 377) for generalized linear models, are in order. In Marx’s algorithm all PLS-directions are used at first to find an estimate of \( \eta \), before iterating on this estimate until convergence. Since we keep the estimation of the baseline hazard and the non-genomic variables outside the PLS regression, it is more convenient for our purpose to start with a single PLS-direction, iterate on the \( \eta, \gamma \) and the baseline hazard \( \Lambda_0 \) estimates until convergence for this direction, and then carry on with the other PLS-directions in the same way (cf. step 2 in Table 2). For each increase in number of PLS-directions we carry with us the estimates of \( \eta, \gamma \) and \( \Lambda_0 \). This modification of Marx’s procedure is mainly algorithmically motivated. Since the first few PLS-directions often will extract most of the information from the data, our approach will more rapidly find reasonable estimates of the baseline hazard and the linear predictors, and therefore give faster convergence.

If we look at the full Cox likelihood (2) with \( x_i^T \beta \) replaced by \( \eta_i \), we see that adding a constant to the \( \eta_i \)'s and subtracting the same constant from the log baseline hazard increments, \( \log \Delta \Lambda_0(t_j) \), leaves the likelihood unchanged. This means there is an indeterminacy in the estimation of these two quantities, which we resolve by the centering in step (d5) of the algorithm. This centering makes \( \sum_{i=1}^{n} \hat{\eta}_i = 0 \), corresponding to the constraint \( \sum_{i=1}^{n} \beta^T \mathbf{x}_i = 0 \) imposed on the linear predictors of an ordinary Cox model with centered covariates.

For model evaluation, which will be described in the next section, an estimate of \( \beta \), together with \( \hat{\gamma} \), is needed. Using the formula provided in Helland (1988), we obtain this by

\[
\hat{\beta} = \hat{W}(\hat{P}^T \hat{W})^{-1} \hat{q},
\]

where \( \hat{W} = (\hat{w}_1, \ldots, \hat{w}_m) \), \( \hat{P} = (\hat{p}_1, \ldots, \hat{p}_m) \) and \( \hat{q} = (\hat{q}_1, \ldots, \hat{q}_m)^T \), using the final estimates of these quantities.

If there are no available additional covariates \( \mathbf{Z} \) available, a special case of the algorithm is applied which simply omits the (e) step in Table 2.
4 Illustrations and comparisons

In this section we apply our PLS Cox method to two breast cancer data sets, one with only gene expression measurements in addition to the patients’ survival times (with censoring information), and one with 6 demographic/clinical variables in conjunction with the genomic and survival data. These data sets are used to compare our method with the original PLS Cox method of Park et al. (2002). The latter data set is also applied to illustrate how our method can be used to include additional non-genomic variables and to estimate survival probabilities. Before we show and discuss the results, we give a brief description of the two data sets.

The van’t Veer data set. The first data set consists of $p = 4919$ gene expression measurements along with survival times and censoring information from $n = 295$ women with breast cancer. The data have been used by van Houwelingen et al. (2006) to illustrate ridge regression for Cox’s model (1), and it contains a subset of the $p = 24885$ gene expressions in the original data from van de Vijver et al. (2002). The data set in van de Vijver et al. (2002) is an extension of the data set with $n = 117$ patients in van’t Veer et al. (2002), whose name will be used as reference for the data set.

The Særlie data set. This data set originates from Særlie et al. (2003) and contains gene expression measurements from 115 women with breast cancer. We use the list of 551 intrinsic genes introduced in Særlie et al. (2003). The data set also include 6 demographic/clinical variables which are complete for 108 of the patients.

4.1 Comparison with the PLS Cox method of Park et al.

We want to compare the performance of our modified PLS Cox method to the original one of Park et al. (2002). As indicated in Section 1, and further discussed in section 5, both methods use the iterated reweighted partial least squares (IRPLS) approach to obtain gene coefficient estimates, but our modified version allows for separate estimation of the baseline hazard rate as well as coefficients for non-genomic variables – if available.

As the method of Park et al. (2002) does not allow for inclusion of non-genomic data, we will only use the gene expression variables in this comparison. This means that we will apply the special case of our method which omits the $\gamma$ estimation step in (e) of our algorithm (Table 2). We first divide a data set into two parts; one training set of about 2/3 of the data for estimation and one test set of the rest of the data for evaluation or testing of the model fitted on the training data. Then we apply 10-fold cross-validation to the
training set to find the optimal number of PLS-directions using Verweij and Houwelingen (1993)'s cross-validated partial log-likelihood (cvpl) criterium, i.e.
\[
    \text{cvpl}(m) = \sum_{k=1}^{10} [\text{pl}(\hat{\beta}^{(-k)}_m) - \text{pl}^{(-k)}(\hat{\beta}^{(-k)}_m)].
\]

Here \(\text{pl}(\beta)\) denotes the log of the partial likelihood (4) for the whole training set and \(\text{pl}^{(-k)}(\beta)\) the same but with the \(k\)-th fold left out, while \(\hat{\beta}^{(-k)}_m\) is the estimate of \(\beta\) using \(m\) PLS directions leaving the \(k\)-th fold out. Given the \(m\) maximizing cvpl\((m)\), we estimate \(\beta\) from the whole set of training patients, and denote this estimate \(\hat{\beta}_{\text{train}}\).

We then assess how well \(\hat{\beta}_{\text{train}}\) is able to predict the outcome of the test data. This will be done in terms of two criteria.

1) **Log rank test.** In the study of cancer one may be interested in assigning the patients to groups based on their prognosis. Here this can be done by calculating the prognostic index (PI):

\[
    \tilde{\eta}_i = x_i^T \hat{\beta}_{\text{train}}
\]

for each test patient \(i\), and then assigning the patients to subgroups according to the ranks of their prognostic indices. The ability of the model to predict the outcome of the test data is then evaluated by the \(P\) value from a log rank test for trend among four equal sized subgroups, see e.g. Klein and Moeschberger (2003, page 217) for the specific formula used.

2) **Deviance.** Another way of evaluating how well the training set estimates are able to predict the outcomes of the test data is to look at the difference in deviance from the null model of the test data using

\[
    -2\{\text{pl}^{(\text{test})(\hat{\beta}_{\text{train}})} - \text{pl}^{(\text{test})}(\mathbf{0})\}.
\]

Here \(\text{pl}^{(\text{test})(\hat{\beta}_{\text{train}})}\) and \(\text{pl}^{(\text{test})}(\mathbf{0})\) are the log of Cox partial likelihood (4) for the test data evaluated at \(\hat{\beta}_{\text{train}}\) and \(\mathbf{0}\). The prediction is good when (13) is small.

Whereas the first criterium just evaluates the correct group assignments, but disregards the ranking within the groups, the second one uses the fit of every single individuals more directly. The first criterium may on the other hand be more relevant for a clinical application of the methods.

After having studied the resulting criteria values from a couple of random 2:1 training/test splits, we found that the results were quite dependent on the particular split. Therefore we performed 50 different training/test partitions for each of the two data sets.
Figure 1: Boxplots comparing our new PLS Cox method with the original method of Park et al. (2002) after performing 50 random splits into training and test sets on the data sets of van’t Veer (top panel) and Sørlie (bottom panel). Left panel shows $P$-values from log rank test for trend among four subgroups based on prognostic index, and right panel shows differences in deviance from the null model. Both criteria are described in Section 4.1.

Figure 1 shows the results in terms of boxplots over the two criteria. The $P$ values from log rank tests are shown in the left panels, and deviance from null model in the right, with the van’t Veer data given in the upper panels and the Sørlie data in the lower. For the van’t Veer data we see that there is not much difference between the two methods in terms of the $P$ values from log rank tests for trend, except that our method shows somewhat less variability. For the deviance we see that our method has slightly lower median and variability. For the Sørlie data the differences are not so clear, but our method has slightly smaller median criteria values. In conclusion, our method gives at least as good predictions as the method of Park et al. (2002) for these data sets.
4.2 Application to data set with additional covariates

To illustrate how non-genomic variables can be included together with gene expression data in our PLS Cox model and how it may be used to estimate survival probabilities we use the data set of Serlie et al. (2003) described above.

The 6 non-genomic variables in the data set are age, lymph node status, estrogen receptor status, tumor thickness, metastasis indicator and grade of disease. Except for age, all of these variables are categorical variables. For each of the variables – also age, which we dichotomize into over and under 55 years – we create dummy variables indicating whether an individual belongs to a category or not. Since the number of individuals belonging to each category could be quite small, we found it necessary to dichotomize all the covariates except from lymph node status and tumor thickness, which were merge into 3 groups.

In a classical setting where predictions are to be done from clinical variables only, one would naturally start out by finding a subset of significant covariates by some variable selection procedure. We do this by backward selection, i.e. by first employing ordinary multivariate Cox regression on all the clinical covariates, then removing the least significant covariate, and proceeding the same way until a covariate set with only significant covariates remains left. This leads to a subset of only two significant covariates: ER status (0=ER negative, 1=ER positive) and grade (0=high grade, 1=low grade), with coefficient estimates 0.84 (se = 0.35) and −0.93 (se = 0.40), respectively.

We now want to investigate whether including the genomic variables along with our clinical variables leads to improved predictions. To do this we use a bit different evaluation procedure than the one described in the previous subsection, in which we, due to the rather small number of patients, employ cross-validation to the whole data set, rather than just the training set, to find the optimal number of PLS-directions for the genomic variables. For most of the folds one or more PLS-directions was picked out indicating that including the genomic variables leads improved predictions. This is confirmed by Table 3 which shows that the log-likelihood, which here, for both covariate models, is summed over the log-likelihood values of the 10 left-out parts in the cross-validation, is increased from −181.6 to −179.0 after inclusion of the genomic variables. We also note that the coefficient estimates change a little when the genomic measurements are included, which is perhaps what one might expect as information on the clinical variables are contained in the gene expressions, but it is quite interesting that the
estimates are not changed more.
In addition to its ability of including non-genomic variables, an advantage of our PLS Cox method is that it provides estimates of the baseline hazard increments, thus enabling estimation of survival curves for given covariate values. Figure 4 shows estimated survival curves for the 12.5%, 37.5%, 62.5% and 87.5% quantiles of PI estimated using the full Sørlie data set (left panel). This is compared with Kaplan Meier survival curves of corresponding subgroups of PI (right panel). Here PI for each individual \( i \) is calculated using \( x_i \) together with the \( \beta \) and \( \gamma \) estimates obtained when leaving out the fold containing individual \( i \) (van Houwelingen et al., 2006). We see from the large spread of the curves of both figures that this model is able to distinguish quite well between the subgroups. The accordance between the model based curves in the left and the empirical curves in the right panels, indicates that the model assumptions (constant and log-linear covariate effects) assumed in the Cox model, are not heavily violated.

5 Discussion
We have presented a modified version of the PLS Cox method of Park et al. (2002), in which the baseline hazard increments, gene effects and effects of additional covariates are estimated in three separate steps. This approach has some advantages over the simultaneous estimation approach of Park et al. (2002). Firstly, it is faster and less memory demanding as it avoids the increase in dimension of the data set required for a simultaneous estimation of the covariate effects and the baseline hazard increments when using the reformulation of the censored survival data problem to a Poisson regression problem. This increase can be quite substantial as the dimension of the rows in the design matrix used in the estimation goes from \( n \) to a maximum of \( n^2/2 \), occurring in the case of no censoring; c.f. the discussion in Park et al.
Figure 2: Estimated survival curves for prognostic index (PI) quantiles vs. Kaplan Meier curves of respective PI subgroups estimated from both clinical and gene expression variables of the Sorli data set. Here R1, R2, R3 and R4 denotes patients with cross-validated PI's (as defined in Section 4.2) between 0-25%, 25-50%, 50-75% and 75-100% PI quantiles, respectively.
(2002). For the van't Veer data set, for instance, the number of rows of the expanded design matrix is 17356, compared to 295 rows for our method, corresponding to about a 59-fold increase. Since the data sets generated by high-throughput-technologies may be expected to become larger as the technologies get more streamlined, the method of Park et al. (2002) might prove to be too memory demanding for many future applications.

Avoiding to expand the design matrix also implies a decrease in computational time. When tested on the full van’t Veer data set using one PLS direction and implementing the line search procedure for optimization of the likelihood described in Section 3 for both methods, our method was about 5 times faster than the method of Park et al. (2002). Fortunately PLS in general is fast, mainly due to the small number of required PLS directions, which by the way is an advantage over the related method principal component regression. In fact, cross-validation applied to the 50 training/test splits of our data sets resulted in the optimal value of $m = 1$ PLS direction for most of the splits. However, this, or some other form of cross-validated approach, has to be employed to find the optimal value of this complexity parameter, a necessity common to all established prediction methods for high-dimensional data, so the computational time is still an important issue also for PLS methods.

In addition to these computational improvements, our method enables inclusion of lower-dimensional demographic/clinical variables. No other PLS methods have focused on how this can be done. Since there for many diseases exist known clinical markers likely to help the predictions, we believe that incorporating both variable sets in many situations would lead to better predictions than using just one of them. Here we have proposed a method for doing this, in which we have tried to take the nature of the variables into account by performing PLS only on the high-dimensional genomic covariates and ordinary regression on the lower-dimensional ones. By this we get PLS directions solely dependent on the genes, not being distorted by clinical/demographic variables of quite different character.

Finally our method has the advantage of giving estimates of the baseline hazard increments, which enables us to estimate survival probabilities for new patients with given covariate values. This is exemplified by the estimated curves in the left panel of Figure 2.
Appendix: The IRLS algorithm for Cox regression

In this appendix we derive the IRLS step (10) for updating the estimate of $\beta$ in Cox’s regression model (1). This is obtained from the Newton-Raphson algorithm for estimating $\beta$ when the baseline hazard increments $\Delta \Lambda_0(t_j)$ and the vector of regression coefficients $\gamma$ are treated as a fixed quantities (offsets). In Section 2 the model is symmetric in the two vectors of regression coefficients, so this will also show that (11) is the IRLS updating step for $\gamma$ when the $\Delta \Lambda_0(t_j)$ and $\beta$ are treated as offsets. We will below without further saying use quantities introduced in Section 2.

Considering the $\Delta \Lambda_0(t_j)$ and $\gamma$ as fixed quantities, the log-likelihood corresponding to (2) is given by

$$l(\beta) = \sum_{i=1}^{n} \left\{ \delta_i \log \Delta \Lambda_0(y_i) + \delta_i \left( \beta^T \mathbf{x}_i + \gamma^T z_i \right) - \Lambda_0(y_i)e^{\beta^T \mathbf{x}_i + \gamma^T z_i} \right\}.$$  

(A.1)

The elements $u_r(\beta) = \partial l(\beta)/\partial \beta_r$ of the vector $\mathbf{U}(\beta) = (u_1(\beta), \ldots, u_p(\beta))^T$ of score functions become:

$$u_r(\beta) = \sum_{i=1}^{n} \left\{ \delta_i x_{ir} - \Lambda_0(y_i)e^{\beta^T \mathbf{x}_i + \gamma^T z_i} \right\} = \sum_{i=1}^{n} \left\{ \delta_i - \mu_i(\beta) \right\} x_{ir}, \quad (A.2)$$

where

$$\mu_i(\beta) = \Lambda_0(y_i)e^{\beta^T \mathbf{x}_i + \gamma^T z_i},$$

while the elements $a_{rs}(\beta) = -\partial u_r(\beta)/\partial \beta_s$ of the information matrix $\mathbf{A}(\beta) = \{a_{rs}(\beta)\}$ may be written

$$a_{rs}(\beta) = \sum_{i=1}^{n} \Lambda_0(y_i)e^{\beta^T \mathbf{x}_i + \gamma^T z_i} x_{ir} x_{is} = \sum_{i=1}^{n} \mu_i(\beta) x_{ir} x_{is}. \quad (A.3)$$

For given values of the $\Delta \Lambda_0(t_j)$ and $\gamma$, we may find the maximizing value of $\beta$ by the Newton-Raphson algorithm. This algorithm updates the current estimate $\hat{\beta}^{(k)}$ of $\beta$ according to the recursion

$$\widehat{\mathbf{A}}^{(k)} \hat{\beta}^{(k+1)} = \widehat{\mathbf{A}}^{(k)} \hat{\beta}^{(k)} + \hat{\mathbf{U}}^{(k)}, \quad (A.4)$$

where $\hat{\mathbf{U}}^{(k)} = \mathbf{U}(\hat{\beta}^{(k)})$ and $\widehat{\mathbf{A}}^{(k)} = \mathbf{A}(\hat{\beta}^{(k)})$ are the current values of the vector of score functions and information matrix, respectively.

The IRLS algorithm is just a convenient reformulation of (A.4). In order to achieve this reformulation, we start out by inserting (A.2) and (A.3) into
the right-hand side of (A.4), to get for its rth component

$$
\left( \hat{A}^{(k)} \hat{\beta}^{(k)} + \hat{U}^{(k)} \right)_r = \sum_{i=1}^{n} \hat{\mu}_{i}^{(k)} x_{ir} \sum_{s=1}^{p} x_{is} \hat{\beta}_{s}^{(k)} + \sum_{i=1}^{n} \left\{ \delta_{i} - \hat{\mu}_{i}^{(k)} \right\} x_{ir}
$$

$$
= \sum_{i=1}^{n} \hat{\mu}_{i}^{(k)} x_{ir} \hat{f}_{i}^{(k)}
$$

(A.5)

with $\hat{\mu}_{i}^{(k)} = \mu_{i}(\hat{\beta}^{(k)})$ and $\hat{f}_{i}^{(k)}$ given by (6). Thus, by (A.5), the right hand side of (A.4) may be written as $X^{T} \hat{V}^{(k)} \hat{f}^{(k)}$. In a similar manner, using (A.3), the left hand side of (A.4) takes the form $X^{T} \hat{V}^{(k)} X \hat{\beta}^{(k+1)}$. Thus the Newton-Raphson algorithm (A.4) may be rewritten as

$$
X^{T} \hat{V}^{(k)} X \hat{\beta}^{(k+1)} = X^{T} \hat{V}^{(k)} \hat{f}^{(k)},
$$

which gives a justification of the IRLS step (10) for updating the estimate of $\beta$.

References


