2010

Combining multiple models with survival data: the PHASE algorithm

Debashis Ghosh, *Penn State University*
Zheng Yuan

Combining multiple models with survival data: the PHASE algorithm

Debashis Ghosh

Departments of Statistics and Public Health Sciences

The Pennsylvania State University

514A Wartik Building

University Park, PA,16802 U.S.A.

ghoshd@psu.edu

and

Zheng Yuan

Eli Lilly and Company

Indianapolis, IN 46285

Abstract

In many scientific studies, one common goal is to develop good prediction rules based on a set of available measurements. This paper proposes a model averaging methodology using proportional hazards regression models to construct new estimators of predicted survival probabilities. A screening step based on an adaptive searching algorithm is used to handle large numbers of covariates. The finite-sample properties of the proposed methodology is assessed using simulation studies. Application of the method to a cancer biomarker study is also given.

1 Introduction

In many medical and scientific studies, it is becoming important to develop prediction rules based on combinations of measurements. One major area of application is in biomarker research. There are a plethora of biomarkers being discovered through genomic and proteomic technologies, for example. This necessitates selecting biomarkers and leads to the statistical problem of variable/model selection. When multiple plausible models are present, the traditional approach is to use a model selection criteria to select a "best" model. This selected model is then used for subsequent inference and prediction. A large amount of literature exists on the topic of model selection (Burnham and Anderson, 2002; Claeskens and Hjort,

2008). Most standard procedures typically ignore the uncertainty in model selection, which leads to poor prediction and diagnostic accuracy on independent data sets.

A promising alternative is model averaging (Hoeting et al, 1999). Work in this area with censored survival data is much more limited. Augustin et al. (2005) have proposed methods by bootstrap re-sampling for combining multiple variables in Cox survival models; their procedure could not handle high-dimensional predictors. A more recent proposal by Zheng et al. (2006) involved combining time-dependent logistic regression models in order to maximize measures of discrimination based on time-dependent receiver operator characteristic curves.

Motivated by the ARMS (Adaptive Regression by Mixing with Screening) method (Yuan and Yang, 2005) and the adaptive model selection method (Shen et al., 2004), we propose a new method for combining Cox models for censored survival data. We term it PHASE (Proportional Hazards Aggregation with adaptive SElection). The paper is organized as follows. In Section 2, we give some background on Cox proportional hazard models. Then we outline the model setup and describe the PHASE algorithm. In section 3, we propose various weighting methods for the algorithm. We study the finite-sample performance of the proposed algorithm through simulation studies in Section 4. In addition, we apply the proposed methodology to tissue microarray data from a prostate cancer study. We conclude with discussion in Section 5.

2 Proposed methodology and inference procedures

2.1 Data and Model setup

Suppose we have p variables available in a study on n individuals. The data are $(Y_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, iid observations from $(Y, \delta, \mathbf{X})$, where Y is the observed time defined as the minimum of failure time T and censoring time C, δ is the indicator of failure ($\delta = 1$) or censoring ($\delta = 0$), and $\mathbf{X}$ is the vector of p multiple biomarkers. Assume that T is independent of C, conditional on $\mathbf{X}$. We consider fitting Cox proportional hazard models of the following form:

$$\lambda(t, \mathbf{X}_i; \beta) = \lambda_0(t) \exp(f(\mathbf{X}_i, \beta)) = \lambda_0(t) \exp(\mathbf{X}_i \beta)$$

where $\mathbf{X}_i\beta$ is the linear predictor in the Cox model and λ_0 is the baseline hazard that is typically left unspecified. There will be $2^p - 1$ models (excluding the trivial intercept-only model) to be considered as candidates. Each model has the form:

$$\lambda_k(t, \mathbf{X}_i; \beta) = \lambda_{0k}(t) \exp(\mathbf{X}_{ki}\beta_k), \quad i = 1, ..., n,$$

where $\mathbf{X_k}$ is a subset of $\mathbf{X}$ and k indexes the model. Estimation of these models is done based on partial likelihood (Cox, 1972).

The goal of model selection is to find a "best" $\mathbf{X_k}_i\beta_k$ that fits the data, while that of model averaging is to combine multiple plausible good models using a weighted combination. The methodology we describe here employs the latter approach.

2.2 PHASE for Cox proportional hazard models

Yang (2001) proposed ARM (adaptive regression by mixing), a method for combining multiple linear regression models. He examined its theoretical convergence properties and empirically demonstrated its adaptive property in nonparametric estimation with a small number of candidate procedures. Yuan and Yang (2005) proposed a modified ARM method with screening in linear regression context, called ARMS.

In this paper, we extend ARMS to the censored survival outcome setting. Originally, Yuan and Yang (2005) used AIC as a screening criterion. When the dimension of p is large, complete searching is not feasible computationally. In order to make our combining algorithm applicable on high-dimensional data and obtain potential better prediction performance for censored survival data, we proposed a new screening method using the idea of adaptive penalty for model selection (Shen et al., 2004). There are four main steps involved. First, half of the sample is used as a training set to estimate the parameters for each model; the other half is used as a test set. Second, we apply the adaptive screening algorithm of Shen et al. (2004) to select a number of most promising candidate models for combining. Third, the response values in the test set are predicted based on the fitted models and the predictions are assessed by comparing the predicted values with the true ones. Finally, the models are weighted according to the performance assessment. The major idea is that models with

good predictive performance will be given larger weights, while those with worse predictive performance will be given smaller weights. For simplicity, assume that the sample size n is even. The following is the proposed PHASE algorithm for Cox proportional hazard models.

1. Split the data into two parts $Z^{(1)} = (Y_i, \delta_i, \mathbf{X}_i), 1 \leq i \leq n/2$ and $Z^{(2)} = (Y_i, \delta_i, \mathbf{X}_i), n/2+ 1 \leq i \leq n$.

2. Suppose that the data $Z^{(1)}$ comprises m_1 uncensored failure times $t_1 < ... < t_{m_1}$. Let j denote the individual failing at t_j, and let $\mathbf{X}_j$ denote the covariate vector for the jth individual. The partial likelihood estimation is applied to $Z^{(1)}$ for each candidate Cox model k. Let $\hat{\beta}_k$ denote the resulting partial likelihood estimate of β_k. The estimate of the baseline survival function $S_0(t)$ is a discrete function with mass points only at $t_1, ..., t_{m_1}$, i.e., $\hat{S}_0(t) = \prod_{j:t_j \leq t}(1 - \hat{\lambda}_j), j = 1, ..., m_1$, where $\hat{\lambda}_j$ is the estimate of the discrete baseline hazard function. Then the survival function estimate is $\hat{S}_k(t; \mathbf{X}, \hat{\beta}) = \hat{S}_{0k}(t)^{\exp(\mathbf{X_k}\hat{\beta}_k)}$ for model k.

3. The log partial likelihood on data $Z^{(1)}$ for model k is

$$\log PL(Z^{(1)}) = \sum_{j=1}^{m_1} \{X_j \hat{\beta}_k - \log[\sum_{i \in R(t_j)} e^{X_i \hat{\beta}_k}]\},$$

   where $R(t_j)$ is the set of subjects at risk at time t_j. Then one computes the adaptive model selection criterion based on the first half data $Z^{(1)}$ as

$$-\log PL_M(Z^{(1)}) + \gamma |M_k|,$$

   where γ is a penalty parameter chosen to range from 0 to 10. The parameter $|M_k|$ is the dimension of the vector of regression coefficients in model k. The optimal model M_γ is found for each γ using the backward stepwise searching proposed by Shen et al. (2004). Then the 100 selected optimal models under different penalty parameters are used as candidates for model combining. Let Γ_s denote the set of these models.

4. Assess the accuracies of the models using the second half of the data $Z^{(2)}$. For each model $k \in \Gamma_s$, compute a model accuracy measurement B_k by a weight. We discuss the issue of what weight to use in the next section.

5. Compute the weight for model k based on B_k in step 4:

$$W_k = \frac{B_k}{\sum_{j\in\Gamma_s} B_j}$$

Note that $\sum_{k\in\Gamma_s} W_k = 1$.

6. Randomly permute the order of the data $N-1$ times and repeat the above 5 steps. Let $W_{k,r}$ denote the weight of model k at the r-th permutation. Then we obtain an average weight $\hat{W}_k = N^{-1}\sum_{r=1}^{N} W_{k,r}$ for each model k over N permutations. N is chosen to be 50 in this work.

7. Let

$$\tilde{S}(t|\mathbf{x}) = \sum_{k\in\Gamma_s} \hat{W}_k \hat{S}_k(t|\mathbf{x})$$

be the final estimator of the survival function given covariates $\mathbf{x}$ at time t.

Although we fit proportional hazard Cox models in the algorithm, our target estimand is in fact the survival function, conditional on covariates. Because of the averaging across multiple models, this decreases the reliance on the true model being of the PH class. The algorithm will adaptively obtain a combined estimator which is a good estimate of the true model even if the true model is not of the proportional hazard form. This is seen theoretically in Theorem 1 of Yang (2001) for the case of linear regression with uncensored outcome. For censored data, it appears harder to derive results of this type.

Our adaptive screening method can remove some very poor models which would affect the performance of the combined estimator. The screening set Γ_s depends on the data splitting. In our numerical examples, every time we permute the data, we will get a different set Γ_s and then work on this different set to combine models. Let $\Gamma \equiv \cup_s \Gamma_s$ denote the set of models we combine over.

The adaptive screening has two advantages over the original AIC screening. First, the adaptive screening does not rely on the accuracy of AIC or any other fixed model selection criteria because it applies an adaptive model selection criterion. The adaptive screening is not as sensitive to the true underlying model as the original AIC screening. Second, due to

the nature of our new adaptive screening, stepwise searching could be used in the adaptive screening procedure and hence can allow for higher dimensional data than the AIC screening procedure of Yuan and Yang (2005). Ideally, the number of models to include in the screening procedure should strike a balance the between the number of models and the variability in the resulting predictions. This is very similar to the bias-variance tradeoff in other areas of statistics. Here, we take 100 models in the screening set by the adaptive screening procedure in this paper.

3 Combining weights in the PHASE algorithm

In Yuan and Yang (2005), only the normal likelihood was used to construct weights for combining results from linear regression models. Augustin et al. (2005) considered a resampling-based algorithm to derive weights. In the Cox proportional hazard models, partial likelihood is frequently used in the inference and is easy to implement in practice. Therefore, we propose three different methods to construct weights used in the PHASE algorithm for combining Cox models based on the partial likelihood. For the definition of the weights, k indexes the model being fitted.

1. *Bernoulli-type partial likelihood weights*

   Suppose that the sample comprises m uncensored failure times $t_1 < ... < t_m$ and ignore for the moment of the case of ties. Let j denote the individual failing at t_j, and let $\mathbf{X}_l$ denote the marker covariate vector for the lth individual. The analog of the normal-likelihood based weights from Yang (2001) to the current setting would be those based on a partial likelihood in the Cox model:

   $$PL_k = \prod_{j=1}^{m} \frac{\exp(\mathbf{X}_{kj}\hat{\beta}_k)}{\sum_{l \in R(t_j)} \exp(\mathbf{X}_{kl}\hat{\beta}_k)},$$

   where $\hat{\beta}_k$ denote the partial likelihood estimate of β_k of Cox model k. The estimate of the baseline survival function $S_0(t)$ is a discrete function with mass points only at $t_1, ..., t_m$, i.e., $\hat{S}_0(t) = \prod_{j:t_j \le t}(1 - \hat{\lambda}_j), j = 1, ..., m$, where $\hat{\lambda}_j$ is the estimate of the discrete baseline hazard function. Then the survival function estimate $\hat{S}_k(t; \mathbf{X}, \hat{\beta}) =$

$\hat{S}_{0k}(t)^{\exp(\mathbf{X_k}\hat{\beta}_k)}$ for model k. The Bernoulli-type likelihood is then defined to be the following for model k:

$$L_k^{BP} = \prod_{i=1}^{n} \hat{S}_k(y_i|\mathbf{x}_i)^{1-\delta_i}(1 - \hat{S}_k(y_i|\mathbf{x}_i))^{\delta_i}.$$

Then weights are constructed to be proportional to L^{BP}:

$$W_k = \frac{L_k^{BP}}{\sum_{j\in\Gamma_s} L_j^{BP}}$$

2. *AIC weights*

   Motivated by Bernoulli-type partial likelihood weights, we use an exponentiated AIC function as the model accuracy measure:

   $$B_k = \exp(-AIC_k^{BP}) = L_k^{BP}\exp(-p_k),$$

   where p_k is the number of regression parameters in model k. Notice that B_k is the product of Bernoulli-type likelihood and a penalty term. Then the corresponding weights are constructed proportional to B_k.

3. *Generalized degrees of freedom (GDF) weights*

   Recently, Ye (1998) and Shen (2004) proposed a concept of the generalized degrees of freedom for a model k. Ye (1998) showed that the GDF can be used as a measure of the complexity or cost of a general modelling procedure. The GDF can be numerically calculated by a Monte Carlo algorithm that we discussed in chapter 2. Motivated by this, we can use the generalized degrees of freedom to replace the number of parameters p_k in partial-AIC weights:

   $$B_k = L_k^{BP}\exp(-GDF_k),$$

   where GDF_k is the generalized degrees of freedom of model k.

All of three above weights have pseudo-Bayesian interpretations as posterior probabilities (Yuan, 2006). They can be viewed as a Bernoulli-type posterior likelihood with certain priors. They are used in steps 4 and 5 of the PHASE algorithm for the assignment of weights. We will compare them in Section 5 in simulation studies and the prostate cancer data.

3. Numerical Results

3.1. Simulation studies

We performed simulation studies in order to assess the finite-sample predictive performance of the proposed algorithm. The procedures were evaluated based on above two criteria: integrated survival difference (ISD) and area under the curve (AUC) evaluated at three time points. The ISD, is defined as the following:

$$ISD = \int_{t=0}^{\tau} (\hat{S}_{ME}(t) - \hat{S}_{KM}(t))^2 dt,$$

where τ is the last observed event time. In the definition of ISD, ME refers generically to the methods that we consider here, while KM denotes the nonparametric Kaplan-Meier estimator for the whole population. Notice that we have suppressed dependence on covariates for the formula above; we present results at the mean values for the covariates. For time-dependent AUC, we follow the procedure of Heagerty et al. (2000) and calculate the AUC at the first, second, and third quartiles of the true failure time distribution. Note that better prediction is associated with lower integrated survival but with higher AUC values. For simplicity, we refer to the partial likelihood, partial AIC, and GDF weights as PHASE-L, PHASE-AIC, PHASE-GDF respectively. We consider two forms of true models: a parametric exponential model and log-normal model. We assume that there are 8 available covariates in all the following simulation studies, unless stated otherwise. Under each form of true models, we further consider two cases for the linear predictor. In the first example, we consider the following true regression model, called Case 1:

$$\mathbf{X}\beta = 1.1X_1 + 1.2X_2$$

In the second example, we consider the following true regression model, called Case 2:

$$\mathbf{X}\beta = 0.1X_1 + 0.2X_2 + 0.3X_3 + 0.4X_4 + 0.5X_5 + X_1X_2$$

Case 1 includes two biomarkers with relatively large coefficients as covariates and Case 2 includes five biomarkers with relatively small coefficients and one pairwise interaction as

covariates. We expect that Case 1 has relatively small uncertainty, while Case 2 has relatively large uncertainty. We will be fitting main effects-only models so that the true model in Case 2 is not even in the space of the regression models being fit.

We first investigated the performance of various methods under the parametric exponential model. We generate a panel of eight biomarkers $\mathbf{X} = (X_1, ..., X_8)$ from multivariate normal with zero mean, unit variance, and correlation 0.3. The failure time is generated from

$$logT = -\log \lambda_0 - \mathbf{X}\beta + W,$$

where W has the extreme value distribution and $\lambda_0 = 1$. The censoring time is generated from a uniform(0, σ_c) where σ_c was chosen to induce about 25% of censoring. Results are given in Tables 1 and 2 for Cases 1 and 2.

From the results for Case 1, in both sample size situations, among the three PHASE methods, both PHASE-AIC and PHASE-GDF methods perform better than the PHASE-L method. This suggests that adding certain penalty terms into the partial likelihood would improve the performance of the partial likelihood. For both sample sizes considered, three PHASE methods have 0-4% smaller survival differences and 0-3% higher AUC values than AIC method. There are only slight differences between the ARMS methods and the AIC selected model. This is mainly because the true model in case 1 is a relatively simple model with only two predictors, which leads to small uncertainty in model selection. Therefore, there is no advantage for using the PHASE method in this case. Fitting a full model in this case includes too many unnecessary parameters, so both PHASE and AIC methods performs better than full model.

From the results for case 2, both the PHASE-AIC and PHASE-GDF methods perform better than the PHASE-L method. The GDF weights have further substantial improvement over AIC weights because GDF is a better model accuracy criterion than AIC (Shen et al., 2004). GDF estimates the predictive loss more accurately than AIC and performs well regardless of whether the true model has a parsimonious representation or not (Shen et.al, 2004). For a sample size of 100, the PHASE-GDF method has 14% smaller survival differences and 9-11% higher AUC values than the AIC selection method. When sample size

is increased to 400, the PHASE-GDF method has about 11% smaller survival difference and 8-9% higher AUC values than the AIC selection method. We observe a greater advantage of PHASE methods over AIC selection method than case 1. This is due to the fact that the model in case 2 has more predictors with smaller coefficients than the model in case 1 which leads to larger uncertainty in model selection procedure. Sample size does affect the performance difference between the PHASE methods and theAIC method in case 2. PHASE performs better than the AIC method when sample size is 100 compared with 400. This is because there is larger uncertainty associated with smaller sample size. Compared with the full model, the AIC selected model does not do better because of the large uncertainty in this case. By contrast, the PHASE methods perform better than the full model.

We now want to assess the robustness of our PHASE methods when the proportional hazard assumption does not hold. In this section, we study the robustness prorperties using a log-normal model as the true model. The biomarkers and censoring time were generated in the same way as exponential model. The failure time is generated from $logT = -\log\lambda_0 - \mathbf{X}'\beta + W$, where W has the standard normal distribution and $\lambda_0 = 1$. Results are given in Tables 3 and 4 for two cases with sample size $n = 100$ and $n = 400$.

Once again, among the three PHASE methods, the PHASE-AIC and PHASE-GDF methods perform better than the PHASE-L method. In Case 1, the PHASE-AIC and PHASE-GDF methods perform the same as the AIC selection method under sample size 100 and fare slightly worse for $n = 400$. In Case 2, the PHASE-GDF method has 8-11% smaller ISD and 5-7% higher AUCs than the AIC-based selection method. The results show a similar pattern to the exponential case. This suggests that our PHASE algorithm has satisfactory robustness properties when the proportional hazards assumption does not hold.

Recently, Zheng et al. (2006) proposed a time-varying logistic regression model for multiple biomarkers in censored survival data. They applied a re-weighted logistic score equation procedure to adjust for the censoring in estimation. They compared the method with other popular methods in survival analysis, such as Cox model, proportional odds model and AFT model. Here we compare our PHASE method with their logistic regression method. The comparison was done in the following simulation setting-up. The censoring time was gener-

ated from a normal with mean μ_c and unit variance, where μ_c was chosen to have approximately 30% censoring. Assume there are five candidate biomarkers: M_1, M_2, M_3, M_4, M_5. The first two markers were generated from a $(50\%, 50\%)$ mixture of $M_1 = Uniform(0, 2)$, $M_2 = Uniform(0, 2) + 5M_1^2$, and $M_1 = Weibull(1, 0.5)$, $M_2 = Weibull(1, 0.5) + 5M_1^2$. In other words, we generated 50% M_1 values from $Uniform(0, 2)$ and the other 50% M_1 values from $Weibull(1, 0.5)$; then we generated 50% M_2 values from $Uniform(0, 2) + 5M_1^2$ and the other 50% M_2 values from $Weibull(1, 0.5) + 5M_1^2$. The remaining three biomarkers were generated independently from standard normal. Then the survival time was generated from a proportional odds model $h(T) = -0.8M_1 - 0.2M_2 + \epsilon$, with $P(\epsilon \leq x) = \frac{e^x}{1+e^x}$ and $h^{-1}(x) = 10\{2\Phi(x/5) - 1\}^3 + 10$. We performed 200 simulations at the sample sizes 200 and 1000. The accuracy measure for comparison is AUC values at $t = t_1, t_2, t_3$, chosen as the 25th, 50th and 75th percentile of the failure time distribution. The results were shown in Tables 5 and 6. the PHASE-AIC and PHASE-GDF methods have consistently 5-7% higher AUC values at the three time points than the time-varying logistic model method for both sample sizes considered. The simulation results suggest the advantage of our PHASE combining method over the method of Zheng et al. (2006).

3.2. Prostate cancer example

In this section, we apply our PHASE method to tissue microarray data from a study in prostate cancer. The biomarkers in the data were measured as continuous protein staining intensities by the Chromavision Medical system. The outcome considers is the time to prostate cancer recurrence. We consider eight biomarkers (ECAD, MIB1,P27,TPD52,BM28, MTA1, AMACR and XIAP) as predictors of the response. We excluded observations with missing values on either response or predictors. This results in $n = 178$ observations.

We made Kaplan-Meier plots to explore the univariate relationship between the biomarkers and time to cancer recurrence. We categorized each of eight biomarkers into 4 groups based on the quartiles of their corresponding empirical distribution. Then we made one stratified Kaplan-Meier plot with four survival curves for each of eight biomarkers. The results are shown in Figure 1. It shows that the survival curves of larger quartile groups are

lower than that of smaller quartile groups for BM28, MTA1, AMACR and XIA. Conversely, the survival curves of larger quartile groups are higher than that of smaller quartile groups for ECAD and MIB1, while there is no clear trend for P27 and TPD52. However, all trends are not significant and we observed that the 4 quartile groups of each biomarker cross between each other quite often. This suggests that using any individual biomarker as a single predictor is not sufficient and may result in poor prediction performance.

We applied our model combining method on the dataset and compared it with the full model, the best model based, and the best univariate model. The data were split into two parts. The first part (119 observations) was used for estimation purposes; the second part (59 observations) was used for validation purposes. In our procedure, we randomly permuted and split the data 1000 times. The accuracy measure we used is the time-specific AUC at three time points, the 25th,50th and 75th percentile of the failure time distribution. The results are shown in Table 7. The results show that the best univariate model has much lower AUC values than any multivariate model. Again, the PHASE-AIC and PHASE-GDF methods perform better than PHASE-Likeli method; and PHASE-GDF method performs much better than the AIC-based selection method. The PHASE-GDF method has higher AUC than the model based on AIC as well as the full model, while the PHASE-L method performs just slightly better than the AIC selection method.

While the focus of the algorithm has been on assessing prediction performance, it is also possible to select biomarkers based on their average weights across models. From the PHASE estimator, we compute $\sum_{k\in\Gamma_s}(\hat{W}_k\hat{\beta}_{kj})$, where we assume $\hat{\beta}_{kj}=0$ if biomarker x_j is not selected in the model k. Thus the average weights for each biomarker is actually $\sum_{k\in\Gamma_s}(\hat{W}_k\hat{\beta}_{kj})$. The results are in Table 8. We see that the rankings tends to be concordant across the univariate and multivariate analysis. From the results, the biomarkers with better univariate analysis prediction results have larger final assigned weights.The two biomarkers, MIB1 and MTA1, were assigned by the final weights 0.77 and 0.72 respectively, which is approximately three times the weight of P27.

4. Discussion

In this article, we have propose a new model averaging method, PHASE, for Cox models with censored survival outcomes and proposed three different types of weights in combining models. We then did simulations studies for comparing our proposed method PHASE with full model and AIC selected model and applied it to a tissue microarray data in prostate cancer study. While being computationally intensive, the results from both simulation examples and the tissue microarray data example showed that PHASE has lower prediction risk and higher AUC than that based on a AIC selected model or the full model. The advantage of PHASE is larger when the underlying true model has more uncertainty for selection procedure and when sample size is smaller. The simulation results also show that the PHASE-GDF method performs better than the other two weights. This is mainly because GDF is a more accurate model assessment criterion than AIC (Ye, 1998).

In PHASE, we employ adaptive screening via an adaptive penalty (Shen et al., 2004). This allow for high-dimensional datasets such as those arising in genomics. Our experience suggests that the adaptive screening is not as sensitive to the true underlying model as the original AIC screening proposed by Yuan and Yang (2005).

In practice, it would be quite simple to use the model combining algorithm. Based on the initial study for validating the panel of biomarkers, one would save the results of the multiple models used for combining and prediction. Given new samples, one could predict the probability of disease at certain time points using the previously saved output.

Appendix: GDF calculation

We first outline the algorithm for calculating generalized degrees of freedom (GDF) when there is no censoring. In that case, we assume the response T is normally distributed with mean μ and variance σ^2. For n individuals, we consider the vector $\mathbf{T} \equiv (T_1, \ldots, T_n)$, which is normally distributed with mean vector whose ith component is μ_i and variance-covariance matrix $\sigma^2 \mathbf{I}_n$. In this setup $\mathbf{I}_n$ is the $n \times n$ identity matrix. A modelling procedure produces fitted values $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_n)$. The GDF for a modelling procedure M are given by

$D(M) = \sum_{i=1}^{n} h_i^M(\mu)$, where

$$h_i^M(\mu) = \frac{dE_\mu[\hat{\mu}_i(\mathbf{T})]}{d\mu_i} = \lim_{\delta \to 0} E_\mu[\frac{\hat{\mu}_i(\mathbf{T} + \delta e_i) - \hat{\mu}_i(\mathbf{T})}{\delta}]$$

$$= \frac{1}{\sigma^2} E[\hat{\mu}_i(\mathbf{T})(t_i - \mu_i)] = \frac{1}{\sigma^2} \mathrm{cov}(\hat{\mu}_i(\mathbf{T}), y_i - \mu_i),$$

where e_i is the ith column of the $n \times n$ identity matrix. We note that the third equality is a result of Stein's identity.

Shen et al. (2004) proposed an idea for estimation of GDF using data perturbation, and we modify that here for the censored data case. In particular, we generate a perturbed version of the censored failure time $Y_i^* = (1 - \tau)Y_i + \tau \tilde{Y}_i, i = 1, ..., n$, where $0 \leq \tau \leq 1$ and $\tilde{Y}_i$ is an independent exponential random variable with hazard function $\Lambda_0 \exp(\mathbf{X}_i \hat{\beta})$, where we took $\Lambda_0 \equiv 2/\tau$. We then refit the PH model to $(Y_i^*, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ and use a slight modification of the algorithm that follows from Theorem 1 of Shen et al. (2004) to estimate the GDF.

References

Augustin, N. H., Sauerbrei, W., Schumacher, M. 2005. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. Statistical Modelling 5, 95 – 118.

Burnham, K. P., Anderson, D. R. 2002. Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach. Springer-Verlag, New York.

Claeskens, G., Hjort, N. L. 2008. Model Selection and Model Averaging. Cambridge University Press.

Cox, D. 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–220.

Heagerty, P., Lumley, T., Pepe, M. 2000. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56, 337–344.

Hoeting, J., Madigan, D., Raftery, A., Volinsky, C. 1999. Bayesian model averaging: a tutorial (with discussion). Statistical Science 14, 382–417.

Shen, X., Huang, H., Ye, J. 2004. Adaptive model selection and assessment for exponential family distributions. *Technometrics* 46, 306–317.

Yang, Y. 2001. Adaptive regression by mixing. Journal of American Statistical Association 96, 574 – 588.

Ye, J. 1998. On measuring and correcting the effects of data mining and model selection. Journal of American Statistical Association 93, 120 – 131.

Yuan, Z. 2006. Ph.D. dissertation, Department of Biostatistics, University of Michigan.

Yuan, Z., Yang, Y. 2005. Combining linear regression models: when and how? Journal of American Statistical Association 100. 1202–1214.

Zheng, Y., Cai, T., Feng, Z. 2006. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. Biometrics 62, 279–287.

Table 1: PHASE results for Case 1 with $n = 400$ under exponential model

Method	ISD	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
ARMS-LIKELI	0.00485	0.794	0.802	0.813
	(0.00013)	(0.007)	(0.007)	(0.006)
ARMS-AIC	0.00475	0.803	0.810	0.822
	(0.00012)	(0.006)	(0.006)	(0.005)
ARMS-GDF	0.00471	0.805	0.812	0.824
	(0.00012)	(0.006)	(0.006)	(0.005)
AIC	0.00473	0.802	0.812	0.811
	(0.00012)	(0.006)	(0.006)	(0.005)
Full	0.00612	0.722	0.735	0.748
	(0.00016)	(0.009)	(0.009)	(0.008)
True	0	0.882	0.899	0.911

Note: Numbers in parentheses are standard errors over 1000 simulations.

Table 2: PHASE results for Case 2 under exponential model

Method	ISD	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
ARMS-LIKELI	0.00223	0.742	0.757	0.773
	(0.00006)	(0.008)	(0.007)	(0.006)
ARMS-AIC	0.00215	0.770	0.776	0.797
	(0.00005)	(0.007)	(0.007)	(0.005)
ARMS-GDF	0.00204	0.790	0.804	0.819
	(0.00005)	(0.007)	(0.007)	(0.005)
AIC	0.00229	0.722	0.739	0.761
	(0.00006)	(0.008)	(0.007)	(0.006)
Full	0.00246	0.703	0.722	0.744
	(0.00007)	(0.010)	(0.008)	(0.007)
True	0	0.867	0.887	0.902

Note: Numbers in parentheses are standard errors over 1000 simulations.

Table 3: PHASE results for Case 1 with $n = 400$ under log-normal model

Method	ISD	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
ARMS-LIKELI	0.0108	0.758	0.754	0.783
	(0.0003)	(0.008)	(0.008)	(0.008)
ARMS-AIC	0.0105	0.769	0.764	0.791
	(0.0003)	(0.007)	(0.006)	(0.007)
ARMS-GDF	0.0104	0.771	0.765	0.794
	(0.0003)	(0.007)	(0.006)	(0.007)
AIC	0.0102	0.778	0.770	0.799
	(0.0003)	(0.007)	(0.006)	(0.007)
Full	0.0130	0.720	0.714	0.748
	(0.0004)	(0.009)	(0.009)	(0.010)
True	0	0.888	0.876	0.909

Note: Numbers in parentheses are standard errors over 1000 simulations.

Table 4: PHASE results for Case 2 with $n = 400$ under log-normal model

Method	ISD	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
ARMS-LIKELI	0.00410	0.736	0.709	0.747
	(0.00011)	(0.008)	(0.007)	(0.008)
ARMS-AIC	0.00395	0.757	0.732	0.768
	(0.00010)	(0.007)	(0.006)	(0.008)
ARMS-GDF	0.00378	0.774	0.750	0.787
	(0.00010)	(0.007)	(0.006)	(0.008)
AIC	0.00409	0.739	0.711	0.750
	(0.00011)	(0.008)	(0.007)	(0.008)
Full	0.00418	0.724	0.699	0.739
	(0.00012)	(0.010)	(0.009)	(0.011)
True	0	0.893	0.863	0.914

Note: Numbers in parentheses are standard errors over 1000 simulations.

Table 5: Estimated AUC from PHASE, weighted logistic regression, and Cox model with $n = 200$ under proportional odds model

	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
PHASE-Likeli	0.86	0.83	0.79
	(0.033)	(0.044)	(0.056)
PHASE-AIC	0.88	0.84	0.80
	(0.032)	(0.044)	(0.056)
PHASE-GDF	0.89	0.85	0.81
	(0.032)	(0.044)	(0.055)
Cox	0.85	0.80	0.76
	(0.042)	(0.070)	(0.089)
Logistic	0.84	0.79	0.76
	(0.066)	(0.067)	(0.091)

Note: Number in parentheses is standard error over 200 simulations.

Table 6: Estimated AUC from PHASE, weighted logistic regression, and Cox model with $n = 1000$ under proportional odds model

	$AUC(t_1)$	$AUC(t_2)$	$AUC(t_3)$
PHASE-Likeli	0.87	0.82	0.79
	(0.011)	(0.014)	(0.019)
PHASE-AIC	0.89	0.84	0.81
	(0.011)	(0.015)	(0.018)
PHASE-GDF	0.90	0.85	0.82
	(0.011)	(0.014)	(0.018)
Cox	0.86	0.80	0.77
	(0.009)	(0.016)	(0.022)
Logistic	0.85	0.80	0.77
	(0.014)	(0.015)	(0.021)

Note: Number in parentheses is standard error over 200 simulations.

Table 7: Estimated AUC for comparing PHASE with AIC and full models in survival analysis on Chromavision data (n=178)

Method	$AUC(t_1)$ risk	$AUC(t_2)$	$AUC(t_3)$
PHASE-Likeli	0.732	0.701	0.739
	(0.013)	(0.012)	(0.013)
PHASE-AIC	0.751	0.722	0.762
	(0.012)	(0.012)	(0.012)
PHASE-GDF	0.776	0.743	0.786
	(0.012)	(0.011)	(0.012)
AIC	0.718	0.689	0.731
	(0.013)	(0.012)	(0.013)
Full	0.702	0.670	0.717
	(0.014)	(0.013)	(0.014)
Univariate*	0.664	0.646	0.685
	(0.013)	(0.013)	(0.014)

Note: Univariate* represents the best univariate model among all univariate models. Numbers in parentheses are standard errors over 1000 permutations.

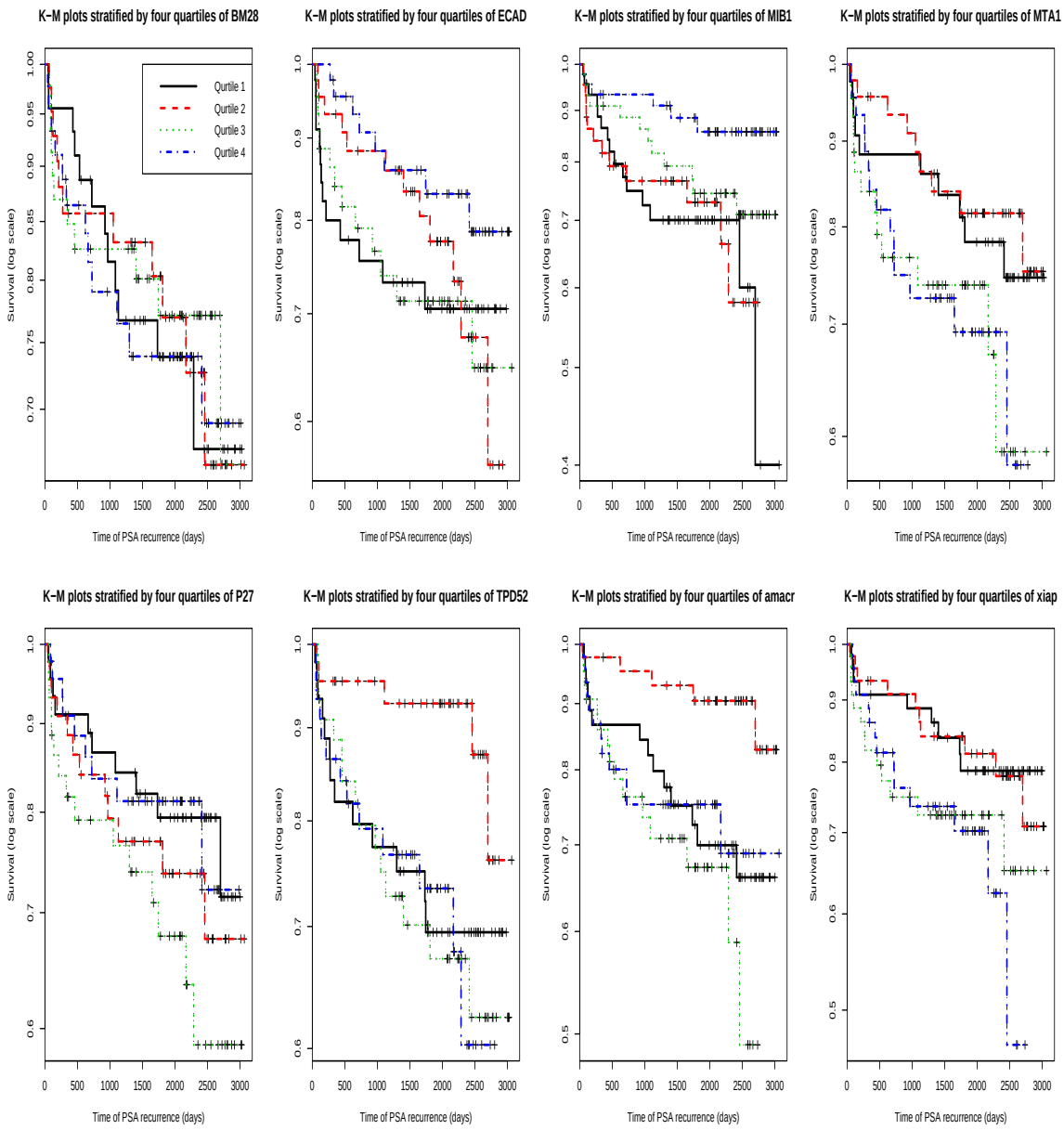

Figure 1: K-M plots stratified by four quartiles of biomarkers. The 1st, 2nd, 3rd and 4th quartile groups are black, red, green and blue lines, respectively.

Table 8: Biomarker weights for prostate cancer data and corresponding univariate prediction performance results: the first row are the final weights assigned to each biomarker predictor by the PHASE algorithm; the second row are prediction results of $AUC(t_2)$ from fitting univariate Cox models of disease status on each biomarker.

Method	BM28	ECAD	MIB1	MTA1	P27	TPD52	AMACR	XIAP
Weights	0.27	0.41	0.77	0.72	0.24	0.35	0.64	0.58
$AUC(t_2)$	0.55	0.58	0.65	0.63	0.54	0.57	0.61	0.60