HHS Public Access

Author manuscript

Commun Stat Theory Methods. Author manuscript; available in PMC 2018 January 09.

Published in final edited form as:

Commun Stat Theory Methods. 2017 ; 46(10): 4791–4808. doi:10.1080/03610926.2015.1085568.

A Simple Method for Deriving the Confidence Regions for the Penalized Cox's Model via the Minimand Perturbation†

Chen-Yen Lin[1] and **Susan Halabi**[2]

[1]Eli Lilly and Company, Indianapolis, IN 46285

[2]Department of Biostatistics and Bioinformatics, Duke University Durham, NC 27710

Abstract

We propose a minimand perturbation method to derive the confidence regions for the regularized estimators for the Cox's proportional hazards model. Although the regularized estimation procedure produces a more stable point estimate, it remains challenging to provide an interval estimator or an analytic variance estimator for the associated point estimate. Based on the sandwich formula, the current variance estimator provides a simple approximation, but its finite sample performance is not entirely satisfactory. Besides, the sandwich formula can only provide variance estimates for the non-zero coefficients. In this article, we present a generic description for the perturbation method and then introduce a computation algorithm using the adaptive least absolute shrinkage and selection operator (LASSO) penalty. Through simulation studies, we demonstrate that our method can better approximate the limiting distribution of the adaptive LASSO estimator and produces more accurate inference compared with the sandwich formula. The simulation results also indicate the possibility of extending the applications to the adaptive elastic-net penalty. We further demonstrate our method using data from a phase III clinical trial in prostate cancer.

Keywords

Cox Model; adaptive LASSO; Confidence Regions; Minimand Perturbation

1 Introduction

Predicting clinical outcomes and identifying prognostic factors are fundamental in medical research. Recent developments in biological and genomic experiments make it possible to collect a vast amount of biomarkers that can be potentially used to build prognostic models of clinical outcomes. Although the increasing availability of biomarkers shed light on promoting knowledge for disease prognosis, it also brings new challenges for statistical modeling. Since the introduction of novel penalized regression methods (Tibshirani, 1996; Fan & Li, 2001), efficient modeling procedures that perform joint parameter estimation and model selection are now the mainstream of modern statistics.

†Supported in part by the National Institutes of Health Grants R01CA155296 and U01CA157703

Corresponding Author: Susan Halabi, susan.halabi@duke.edu.

Researchers are particularly interested in relating a set of baseline covariates to patient's survival time within the context of phase III clinical trials. This task is commonly performed by the Cox's proportional hazards model (Cox, 1972). In the proportional hazards model, the dependency of the survival time t on the predictors $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is characterized by a hazard function of the form

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp[\boldsymbol{x}^T\boldsymbol{\beta}_0], \quad (1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is a p-dimensional unknown regression parameters, commonly called the log-hazard ratios. In the semi-parametric model framework, the primary objective is to make inferences on the unknown regression parameters.

In addition to estimating the unknown regression parameters, identifying a subset of prognostic covariates is another important, yet challenging, inferential topic. For this purpose, numerous variable selection tools that were originally designed for linear regression, such as stepwise selection, hypothesis testing procedure, and Bayesian model selection (Ibrahim, Chen, & MacEachern, 1999), have been extended to the Cox's model. However, their theoretical properties were not well-studied (Fan & Li, 2002). Recently, a class of penalized procedures, including but not limited to the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1997; Park & Hastie, 2007), the adaptive LASSO (Zhang & Lu, 2007), the smoothly clipped absolute deviation (SCAD) (Fan & Li, 2002; Bradic, Fan, & Jiang, 2011), and the elastic net (Wu, 2012), were proposed for the survival endpoint. These procedures are appealing for their shrinkage property that produces sparse point estimates and therefore achieves joint selection and estimation.

The theoretical properties of the penalized estimators have been extensively studied in the literature (Fan & Li, 2002; Zhang & Lu, 2007; Bradic et al., 2011). Although these penalized methods are effective for variable estimation and selection, their limiting distributions are complicated and difficult to approximate analytically (Knight & Fu, 2000; Chatterjee & Lahiri, 2011; Minnier, Tian, & Cai, 2011). As a result, it is crucial to have an alternative numerical approach that can be implemented to approximate the distributions of the regularized estimators.

In practice, it is expected to report a standard error estimate to the proposed estimator or provide an interval estimator. Both standard error and interval estimators aim to quantify the uncertainty of the proposed estimator. In the absence of a valid uncertainty assessment, the point estimate alone does not provide enough information for clinical decision making. Tibshirani (1997) initially suggested a ridge-type of approximated standard error formula for the LASSO estimator. It has, however, the drawback of providing zero standard error estimates for the zero estimated coefficients. Fan and Li (2002) and Zhang and Lu (2007) proposed sandwich formulas for estimating the variance. Their formulas, however, also suffered from a similar limitation in that it only provided variance estimates for the non-zero coefficients. In a simulation study, Zhang and Lu (2007) demonstrated that there was a

discrepancy between the sandwich-based standard error and the Monte Carlo-based standard error.

To alleviate the limitations of sandwich formulas, various numerical methods have been proposed in the context of linear regression to assess the variances of the penalized estimators. Knight and Fu (2000) considered a residual-based bootstrap method for the LASSO estimator in linear regression. Chatterjee and Lahiri (2010) studied the residual-based bootstrap method and showed that the asymptotic distribution of the bootstrapped LASSO estimator is a random measure. Hence, the bootstrap estimator is inconsistent whenever one or more components of the regression parameter is zero. To overcome the difficulty of the traditional bootstrap, Chatterjee and Lahiri (2010) recommended a truncated LASSO estimator whose distribution can be approximated using a residual-based bootstrap method. Recently, Chatterjee and Lahiri (2011) proposed a modified bootstrap method which provides a valid approximation.

Minnier et al. (2011) introduced a resampling method based on minimand perturbation (Jin, Ying, & Wei, 2001) to derive a confidence region for regularized procedures possessing the *oracle* properties, such as the adaptive LASSO (Zou, 2006) and the SCAD (Fan & Li, 2001). In most clinical trials in cancer, the primary outcome is time-to-event endpoint. However, this resampling approach has not been applied to time-to-event outcomes. This could be due to the fact that the notion of residual in the Cox's model is not compatible with that in the linear regression model. We adapt the minimand perturbation method to construct confidence regions for the Cox model, regularized by the LASSO penalty and the adaptive elastic net-penalty. To our knowledge, no one has utilized this approach to the Cox's model.

This article is organized in the following manner. We first review the penalized estimation in the Cox's model and then introduce our perturbation method in Section 2. We present the computation algorithm and the construction of confidence regions in Section 2. We evaluate the proposed method using simulations and real data and present the results in Sections 3 and 4, respectively and we provide some concluding remarks in Section 5.

2 Minimand Perturbation Procedure

2.1 Penalized Partial Likelihood Estimation and Oracle Properties

Consider an analysis with time-to-event outcome, we denote the observed triplet as $\{(t_i, \delta_i, \boldsymbol{x}_i): i = 1, \ldots, n\}$, where t_i is the survival time if $\delta_i = 1$, and censored time if $\delta_i = 0$, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional predictors. Under the proportional hazards model framework in (1), the estimation of the regression parameters is commonly carried out by minimizing the minus log-partial likelihood function

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} n^{-1}\mathscr{L}(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n}\delta_i\left[\boldsymbol{x}_i^T\boldsymbol{\beta} - \log\sum_{j\in R_i}\exp[\boldsymbol{x}_j^T\boldsymbol{\beta}]\right], \quad (2)$$

where $R_i = \{j: t_j \geq t_i\}$ is the risk set just before time t_i. We refer to the minimizer of (2), $\tilde{\beta}$, as the maximum partial likelihood estimator (MPLE).

The MPLE will not produce exact zero estimates in general, even though the true regression parameters are sparse. To produce exact zero estimates and thus achieve joint estimation and selection, we consider adding a penalty term into (2) and solving the penalized log-partial likelihood function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, n^{-1}\mathscr{L}(\boldsymbol{\beta}) + \sum_{j=1}^{p} J_{\lambda_n}(|\beta_j|), \quad (3)$$

where J_{λ_n} is a non-negative penalty function and λ_n is a non-negative regularization parameter that governs model sparsity. If λ_n is large enough, the penalty function $J_{\lambda_n}(\cdot)$ would shrink some small coefficients to exact zeros. The shrinkage property is appealing in terms of identifying the sparse structure, but it also biases the point estimates toward zeros, especially for the large β_{0j}. Hence, an ideal penalty function should be able to produce sparse estimates while keeping the large β_{0j} intact. The reader is referred to Fan and Li (2001) for more discussion on the properties of penalty functions.

In this article, we consider the adaptive LASSO (ALASSO) penalty that has the form $J_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|/|\tilde{\beta}_j|$. The inverse absolute estimates $1/|\tilde{\beta}_j|$ serves as data-driven weights to avoid over-penalizing large parameters and hence give rise to the name adaptive LASSO. The choice of adaptive weights is crucial for optimizing the performance of the ALASSO estimator. We propose to adopt the inverse of the absolute MPLE as adaptive weights for convenience because the MPLE is consistent and asymptotically normal (Tsiatis, 1981) and its magnitude can better represent the relative importance of the predictors. We are particularly interested in the ALASSO penalty for its wide popularity in the statistical literature and its oracle properties (Zhang & Lu, 2007).

According to Fan and Li (2001), an estimation procedure is said to enjoy the oracle properties if: (1) it identifies the correct sparse representation, and (2) the root-n consistent estimator is asymptotically normal for the non-zero parameters as sample size goes to infinity. Moreover, the asymptotic precision matrix is the same as the Fisher information knowing the true sparse representation. Although theoretically sound, the oracle properties rarely hold in finite sample and the inference based on the asymptotic result is unreliable (Minnier et al., 2011; Chatterjee & Lahiri, 2011).

2.2 Perturbed Penalized Partial Likelihood

To approximate the asymptotic distribution of $\hat{\beta}$ more accurately, we consider a minimand perturbation method of Jin et al. (2001). The minimand perturbation was originally proposed to make inference on estimators whose covariance matrix may be difficult to estimate analytically.

The notion of minimand perturbation is simply assigning a random weight to each likelihood contribution. Let $\{G_i: i = 1, \ldots, n\}$ be a set of *i.i.d.* non-negative random sample drawn from an underlying distribution with mean μ and variance $K^2\mu^2$. The stochastic perturbation of $\mathscr{L}(\boldsymbol{\beta})$ is given by

$$n^{-1}\mathscr{L}^*(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n} G_i\delta_i \left[\boldsymbol{x}_i^T\boldsymbol{\beta} - \log\sum_{j\in R_i}\exp[\boldsymbol{x}_j^T\boldsymbol{\beta}]\right]. \quad (4)$$

Let $\tilde{\boldsymbol{\beta}}^*$ be the minimizer of (4), then under suitable regularity conditions, Jin et al. (2001) showed that the distribution of $\tilde{\boldsymbol{\beta}}^*$ can be used to approximate the distribution of $\tilde{\boldsymbol{\beta}}$. In practice, the distribution of $\tilde{\boldsymbol{\beta}}^*$ is empirically estimated by repeatedly solving (4), say B time, using different random sample $\{G_i: i = 1, \ldots, n\}$. Then the inference on $\tilde{\boldsymbol{\beta}}$ is made based on the empirical distribution of $\left\{\tilde{\boldsymbol{\beta}}_b^*, b=1,\ldots,B\right\}$. For instance, the covariance matrix of $\tilde{\boldsymbol{\beta}}$ can be estimated by the sample covariance matrix.

Following the same analogy, for the b-th realization of $\{G_i: i = 1, \ldots, n\}$, we stochastically perturb the penalized objective function in (3)

$$-n^{-1}\sum_{i=1}^{n} G_i\delta_i \left[\boldsymbol{x}_i^T\boldsymbol{\beta} - \log\sum_{j\in R_i}\exp[\boldsymbol{x}_j^T\boldsymbol{\beta}]\right] + \lambda_n^*\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|, \quad (5)$$

where λ_n^* is another smoothing parameter and we defer the tuning procedure to the next section. Let $\hat{\boldsymbol{\beta}}_b^*$ be the minimizer of (5), we propose to use the empirical distribution of $\left\{\hat{\boldsymbol{\beta}}_b^*, b=1,\ldots,B\right\}$ to approximate the distribution of $\hat{\boldsymbol{\beta}}$.

In order for the asymptotic oracle properties to hold and the perturbation method to work, Minnier et al. (2011) established three key regularity conditions. For the Cox's model, we argue that all the required regularity conditions hold. First, the convexity of the log-partial likelihood function ensures the unique solution at the true value as well as the positive definiteness of the Hessian matrix. Kosorok (2008) established that the log-partial likelihood is Glivenko-Cantelli. Finally, the log-partial likelihood is continuous and differentiable (Tsiatis, 1981) and its derivative meet the properties described in the regularity condition.

2.3 Computation

The optimization problem in (5) is essentially the same as the standard ALASSO optimization problem. The random perturbation will neither affect the convexity of the objective function nor the computation difficulty and hence the optimization problem can be solved efficiently. To solve the ALASSO optimization problem for the Cox's model, Simon, Friedman, Hastie, and Tibshirani (2011) proposed an iteratively re-weighted least squares method combined with the state-of-art coordinate descent algorithm (Friedman, Hastie,

Höfling, & Tibshirani, 2007). In this article, we adapt their algorithm to solve the optimization problem in (5).

Let $\boldsymbol{\eta}=(\eta_1 \ \dots \ \eta_n)^T=(\boldsymbol{x}_1^T\boldsymbol{\beta} \ \dots \ \boldsymbol{x}_n^T\boldsymbol{\beta})^T$, and denote $\dot{\mathscr{L}}^*$ and $\ddot{\mathscr{L}}^*$ as the gradient vector and Hessian matrix of $\mathscr{L}^*$ with respect to $\boldsymbol{\eta}$, respectively. Let $\breve{\boldsymbol{\beta}}$ be a proper initial value and $\breve{\boldsymbol{\eta}}=\boldsymbol{X}\breve{\boldsymbol{\beta}}$, by using the same argument in Simon et al. (2011), we can show that the perturbed log-partial likelihood function can be approximated by a second-order Taylor expansion. In particular,

$$n^{-1}\mathscr{L}^*(\boldsymbol{\beta}) \approx \frac{1}{2n}(\boldsymbol{z}(\breve{\boldsymbol{\eta}})-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}(\breve{\boldsymbol{\eta}})(\boldsymbol{z}(\breve{\boldsymbol{\eta}})-\boldsymbol{X}\boldsymbol{\beta})+C, \quad (6)$$

where $\boldsymbol{z}(\breve{\boldsymbol{\eta}})=\breve{\boldsymbol{\eta}}-\boldsymbol{V}^{-1}\dot{\mathscr{L}}^*(\breve{\boldsymbol{\eta}})$ is a pseudo response vector, $\boldsymbol{V}=\mathrm{diag}(\ddot{\mathscr{L}}^*(\breve{\boldsymbol{\eta}})_{1,1} \ \dots \ \ddot{\mathscr{L}}^*(\breve{\boldsymbol{\eta}})n,n)$ is a diagonal matrix containing diagonal elements of $\ddot{\mathscr{L}}^*(\breve{\boldsymbol{\eta}})$, and C is some constant that does not depend on $\boldsymbol{\beta}$. We summarize the algorithm of solving $\hat{\boldsymbol{\beta}}^*$ into following steps

1. Initialize $\breve{\boldsymbol{\beta}}$ and set $\breve{\boldsymbol{\eta}}=\boldsymbol{X}\breve{\boldsymbol{\beta}}$.
2. Compute $\dot{\mathscr{L}}^*(\breve{\boldsymbol{\eta}})$, $\boldsymbol{V}(\breve{\boldsymbol{\eta}})$ and $\boldsymbol{z}(\breve{\boldsymbol{\eta}})$.
3. Find $\hat{\boldsymbol{\beta}}^*$ that minimizes

$$n^{-1}\sum_{i=1}^{n}v_{ii}(z(\breve{\eta})_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\lambda_n^*\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|. \quad (7)$$

4. Set $\breve{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^*$ and $\breve{\boldsymbol{\eta}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}^*$.
5. Repeat step 2–4 until $\hat{\boldsymbol{\beta}}^*$ converge.

In step 3, we drop the constant C and absorb the fraction 1/2 into the regularization parameter λ_n^*. The above algorithm requires repeatedly minimizing a weighted squared loss plus an ALASSO penalty. We use the `R` package `glmnet` (Friedman, Hastie, & Tibshirani, 2012) that implements the coordinate descent algorithm to solve the minimization problem. Finally, at each iteration, we need to update the current weight and the pseudo response using the formulas

$$v_{kk}=\ddot{\mathscr{L}}^*(\breve{\eta})_{kk}=\sum_{l\in C_k}G_l\delta_l\frac{e^{\breve{\eta}_k}}{\sum_{j\in R_l}e^{\breve{\eta}_j}}\left[1-\frac{e^{\breve{\eta}_k}}{\sum_{j\in R_l}e^{\breve{\eta}_j}}\right], \ k=1,\dots,n, \quad (8)$$

and

$$z(\breve{\eta})_k=\breve{\eta}_k-\frac{\dot{\mathscr{L}}^*(\breve{\eta})_k}{\ddot{\mathscr{L}}^*(\breve{\eta})_k}=\breve{\eta}_k+\frac{1}{v_{kk}}\left[G_k\delta_k-\sum_{l\in C_k}G_l\delta_l\frac{e^{\breve{\eta}_k}}{\sum_{j\in R_l}e^{\breve{\eta}_j}}\right],k=1,\ldots,n, \quad (9)$$

where $C_k=\{i: t_i\le t_k\}$.

2.4 Extension to the Adaptive Elastic-Net Penalty

While a sandwich-type of standard error estimator have been proposed for the ALASSO, a standard error estimator remains void for the adaptive elastic-net penalty (Zou & Zhang, 2009). We experiment by extending the application of the proposed method to the adaptive elastic-net penalty. With some modification, the adaptive elastic-net penalty has the form

$$J_{\lambda_n}^{\mathrm{EN}}(|\beta_j|)=\lambda_n\left[\frac{1-\alpha}{2}|\beta_j|^2+\alpha w_j|\beta_j|\right], \quad (10)$$

where w_j's are the adaptive weights. For consistency with the ALASSO, we let $w_j=|\tilde{\beta}_j|^{-1}$.

When appending the adaptive elastic-net penalty to the minus log-partial likelihood, we denote the minimizer as $\hat{\boldsymbol{\beta}}_{\mathrm{EN}}$, that is,

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, n^{-1}\mathscr{L}(\boldsymbol{\beta})+\sum_{j=1}^{p}J_{\lambda_n}^{\mathrm{EN}}(|\beta_j|). \quad (11)$$

In the Cox's model context, Wu (2012) modified the least angle regression algorithm (Efron, Hastie, Johnstone, & Tibshirani, 2004) and proposed a solution path for solving the objective function of (11). The optimization of (11) can also be carried out using the iteratively-reweighted least squares method described in Section 2.3 by changing the penalty term in (7) to an adaptive elastic-net penalty. To simplify exposition, we continue utilizing the iteratively-reweighted process to solve the regularized Cox's model with an adaptive elastic-net penalty.

2.5 Parameter Tuning and Confidence Regions

In order for the ALASSO estimator to enjoy the oracle properties, the regularization parameter needs to be chosen such that $\sqrt{n}\lambda_n\to 0$ and $n\lambda_n\to\infty$ (Zhang & Lu, 2007). For practical use, we may choose λ_n and λ_n^* based on the information criteria, such as the AIC or the BIC, or cross-validation. In this study, we consider an AIC-type of criterion to select the regularization parameters in (3) and (5). For a given λ, we denote

$$\mathrm{AIC}(\lambda_n) = 2\mathscr{L}(\hat{\boldsymbol{\beta}}(\lambda)) + \min(n^{0.1}, 2)\widehat{df}(\lambda_n), \quad (12)$$

and

$$\mathrm{AIC}^*(\lambda_n^*) = 2\mathscr{L}^*(\hat{\boldsymbol{\beta}}^*(\lambda_n^*)) + \min(n^{0.1}, 2)\widehat{df}(\lambda_n^*), \quad (13)$$

where $\widehat{df}(\cdot)$ is the number of non-zero elements in the parameter estimates. We choose the AIC-type of penalty factor min($n^{0.1}$, 2), instead of the BIC-type of penalty factor log(n), to avoid excess shrinkages on the parameter estimates. From our experience, when the BIC criterion is used, the empirical coverage rate is much smaller than the nominal rate as a result of the biased parameter estimates and deflated standard error. Although cross-validation approach also tends to choose a less amount of shrinkage. we adopt the AIC-type of penalty for computational consideration.

In considering the absence of a well-defined degree of freedom estimate for the adaptive elastic-net penalty, the information criteria cannot be applied for tuning the regularization parameters in the adaptive elastic-net penalty. Instead, cross-validation will be used to identify the optimal pair of (λ_n, α) that minimizes the cross-validated model deviance. In the subsequent computation example, we use 10-fold cross-validation to compute the model deviance.

We construct three types of (1 − α)100% confidence regions (CR). The first type is based on the normal approximation and the second one is based on the empirical quantiles. In particular, we estimate the standard error for the j-th coefficient by $\hat{\sigma}_j = \sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\beta}_{jb}^* - \hat{\beta}_j)^2}$, and then the first CR for β_{0j}, CR^{N}, is given by $\hat{\beta}_j \pm \Phi^{-1}(\alpha/2)\hat{\sigma}_j$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal distribution function. The second type of CR, CR^{Q}, is constructed by taking the empirical (α/2)100th and (1 − α/2)100th quantiles as the lower and upper bounds. The third type of CR is also based on the normal approximation but the standard error is estimated by the sandwich formula given in equation (8) of Zhang and Lu (2007). It is noteworthy that the sandwich formula can only provide standard error estimate for non-zero coefficients. For the zero coefficients, the confidence region reduces to a single point.

3 Simulation Study

3.1 Simulation Setup

We conduct simulation studies to examine the performance of the proposed perturbation method. Three simulation examples are considered.

> **Example 1:** In the first example, the predictors x_{ij}, i = 1, …, n, j = 1, …, p, are marginally generated from the standard normal distribution and their pairwise

correlation is $\text{cor}(x_{ij}, x_{ik}) = 0.3^{|j-k|}$. We generate the survival time from an exponential distribution with hazard function

$$\lambda(t_i|\boldsymbol{x}_i)=1 \cdot \exp[\boldsymbol{x}_i^T \boldsymbol{\beta}_0], \quad (14)$$

where the true regression parameters in this example are $\boldsymbol{\beta}_0 = (1, -1, 0, 0, 1, 0, \cdots)^T$. We independently generate a censored time from $U(0, a_1)$, where a_1 is chosen such that 30% of the survival times are censored.

Example 1a: We consider a high-dimensional variation of Example 1 by setting the number of predictors to $p = 1000$. When the number of predictor is greater than the sample size, the MPLE does not exist due to singularity. A principled sure independent screening of Zhao and Li (2012) is utilized to reduce the number of predictor down to $\lfloor n/\log(n) \rfloor$. After the initial screening stage, the remaining steps will follow the same analytical procedure as described above.

Example 2: In the second example, we consider a different correlation structure between the predictors and a weaker signal strength. The predictors are still marginally generated from a standard normal distribution, but their pairwise correlation follows a compound symmetry structure, $\text{cor}(x_{ij}, x_{ik}) = 0.2$. The survival time is generated from a Weibull distribution with hazard function

$$\lambda(t_i|\boldsymbol{x}_i)=\gamma t_i^{\gamma-1}\exp[\gamma \boldsymbol{x}_i^T \boldsymbol{\beta}_0], \quad (15)$$

where the shape parameter $\gamma = 0.5$. The true regression parameters in this Example are $\boldsymbol{\beta}_0 = (1, -1, 0, 0, 0.75, -0.75, 0, \cdots)^T$. Similar to Example 1, an independent censoring time is generated from $U(0, a_2)$ and the constant a_2 is chosen such that 15% of the survival times are censored.

Example 3: The third simulation example examines the practicability of extending the proposed method to the adaptive elastic-net penalty. In this example, we keep the same simulation scenario as that in Example 1 but change the pairwise correlation to $0.5^{|j-k|}$ and control the censoring rate at 20%. Considering the absence of well-accepted variance estimator for the adaptive elastic-net penalty, we will only present the performance of the proposed method and compare it with the MPLE.

We consider three sample sizes $n = \{200, 400, 800\}$ throughout the three examples and two number of predictors $p = \{10, 20\}$, except for Example 1a in which the number of predictors is $p = 1000$, combinations. We perturb the minimand $B = 300$ times and the random variables G_i's are taken from a Gamma distribution with parameters (1,1), which is also used in Jin et al. (2001). The process is repeated 500 times for each scenario and we summarize the average performance by the empirical coverage probabilities and length.

Throughout the simulation and real data analysis, we construct the CRs using the nominal rate of 90% and demonstrate that the empirical probabilities of covering the true parameters. In addition, we compare our method with two other methods: the oracle estimator and the

MPLE. Because the oracle method knows the true sparse representation *a priori* and only needs to estimate the non-zero parameters, it mainly serves as a reference. The CRs for both the oracle and the MPLE are based on the normal approximation.

3.2 Simulation Results

We first consider the case when there are 10 predictors and summarize the simulation results in Table 1. The results presented in Table 1 are consistent with the established theoretical properties. Since the oracle and the MPLE are both consistent and asymptotically normal, their empirical coverage rates should be very close to the nominal rate. As expected, as presented in columns 2 and 4, all the empirical coverage rates are fairly close to 90%. The MPLE is less efficient and thus has wider CRs than the oracle estimator.

As the sample size gets larger, the ALASSO estimator should identify the prominent predictors and efficiently estimate their associated parameters. In the simulation result, as the sample size increases, we observe that the empirical coverage rates for the non-zero parameters tend to the nominal rate and the lengths also become closer to those of the oracle estimator. In addition, for the zero parameters, the empirical coverage rates are all higher than the nominal rate and tending to one as the sample size increases, suggesting that the ALASSO estimator is asymptotically as efficient as the oracle estimator.

Both types of CRs that we proposed for the ALASSO perform well. Because the AIC-type of selection criterion does not provide enough shrinkage, the ALASSO point estimates are not close enough to zero and hence CR^N covers the zero parameters less frequently than the asymmetry quantile-based CR. Both of the symmetric CRs, CR^S and CR^N, provide comparable coverage probabilities. The sandwich formula produces smaller variances estimates and thus narrower confidence intervals are observed across the zero and the non-zero coefficients, respectively. It is noteworthy that since the sandwich formula is applicable for the non-zero estimates, the comparison between CR^N and CR^S to the non-zero coefficients will be more informative.

We illustrate the simulation results for the 20 predictors case in Figure 1. As expected, the empirical coverage rates for the oracle and the MPLE methods are all close to 90%. Resulting from the shrinkage property, the adaptive LASSO point estimates are biased toward zeros and therefore all of the three CRs have difficulty covering the non-zero parameters when the sample size is relatively small. As the sample size increases, the coverage rates converge quickly to the nominal rate and one for the non-zero and the zero parameters, respectively.

As presented in Figure 1, it is clear that the MPLE has the same variability across all unknown parameters. In contrast, the ALASSO estimator has very distinct variabilities across the zero and the non-zero parameters. For the zero parameters, the variabilities are much smaller and the coverage probabilities tend to one. As for the non-zero parameters, the coverage rate and the length are similar to those of the oracle estimator, reflecting the superior performance of the ALASSO estimator and its oracle properties.

To further examine the accuracy of the proposed variance estimator, we compare the standard errors with their estimates. Table 2 summarizes the mean of the estimated standard errors and the sample standard error from the Monte Carlo simulations. For all sample size and the number of predictors combinations, there is minor difference between the estimated standard errors and the sample standard errors. However, the difference vanishes as sample size increases. We again observe that the sandwich formula under-estimates the true variability in Table 2.

In Example 1a, a screening procedure is utilized to reduce the number of predictors to a more manageable scale. The frequency of reserving all predictors with non-zero coefficients, also known as sure screening, is presented in Table 3. Also presented in Table 3 is the required screening size in order to keep all predictors with non-zero coefficients. Since the predictors with non-zero coefficients may not be reserved after the screening stage, the simulation results shown in Table 4 are based on the repetitions when they were reserved after screening. In order to simplify the presentation, Table 4 only presents the empirical coverage and length for predictors with non-zero parameters. When the sample size is 200, only 15.2% of the times all the prominent predictors were reserved after the screening stage. Thus, making the downstream inference unreliable as demonstrated by the under-covered empirical probability. When the sample size increases to 800, 120 predictors will be reserved during the screening, the sure screening rate is almost 100%, and the performance is closer to what was displayed in Table 1. However, as a result of a greater number of predictors, the results shown in Table 4 are still not as efficient compared with those presented in Table 1.

We summarize the simulation results for Example 2 with 10 predictors in Table 5. The simulation results are similar to those in the first example. The coverage rates for the oracle and the MPLE remain fairly close to the nominal rate. CR^N and CR^Q produce the highest coverage for the non-zero and the zero parameters, respectively. The length of CR^Q is however shorter and therefore costs some coverage rates for the non-zero parameters in this example. We also observe that the length of CR^S is almost identical to that of the oracle. Moreover, the coverage rates for the non-zero parameters are consistently lower than those of CR^N, implying that the sandwich formula indeed under-estimates the standard error. We also compare the mean of the estimated standard errors and the sample standard error to assess the accuracy. The results are similar to what was observed in example 1 and therefore will not be further discussed.

We demonstrate the simulation result for the 20 predictors scenario using Figure 2. Due to the stronger signals, the first two non-zero parameters are estimated well across different sample sizes. As the sample size increases, the coverage rates for the 5- and 6-th parameters gradually catch up and close to the nominal rate. In this case, as an indication of under-estimate the variability, we still observe the coverage rates for CR^S are consistently lower than those of CR^N. In conclusion, our perturbation-based CRs still perform well even though the signal strength are weaker and the correlation between predictors are more complicated.

The simulation results for the third example is presented in Table 6. These results indicate promising potential of applying the proposed method, particularly the CR based on the normal approximation, to the adaptive elastic-net penalty. When the sample size is 200, the

CRs based on normal approximation have coverage rates of 85.2, 84.4 and 89.4 for the 1st, 2nd and 5th predictors, respectively. Although slightly under the nominal rate when the sample size is 200, the coverage rates are sufficiently close to the nominal rate with sample size of 400. The quantile-based CR, however, does not provide sufficient coverage even with sample size of 800. We conjecture that the under-coverage is due to the additional ridge penalty as the perturbed point estimates are further shrunk toward zero and hence the quantiles of the perturbed point estimate are not wide enough to cover the true parameter.

4 Real Data Analysis

Our goal is to develop a prognostic model of overall survival using data from a phase III trial of metastatic castration-resistant prostate cancer (CRPC)(Kelly et al., 2012). The trial, CALGB 90401, enrolled 1,050 men with CRPC and part of the data was reserved for testing purpose. We restrict our attention to 536 patients in the training set without missing predictors. The list of 16 baseline variables and their summary statistics are presented in Table 7. Among 536 patients with complete observations, 486 of them died before the end of the follow-up period.

We first fit an unpenalized Cox's model using the `coxph` function in `R`. The unpenalized estimates are used to construct the adaptive weights for ALASSO. Then we fit an ALASSO model using the `glmnet` function to select important baseline covariates and estimate their hazard ratios. Following the same analysis procedure, the regularization parameter is selected by the AIC and the ALASSO estimates 13 non-zero coefficients. We then construct CRs using the sandwich formula and our proposed method. In this analysis, we perturb the minimand $B = 500$ times. The ALASSO point estimates and their associated CRs are presented in Figure 4.

By the equivalence between hypothesis test and confidence interval, we also conduct a test for the null hypothesis H_0: $\beta_{0j} = 0$, $j = 1, \ldots, 17$. The null hypothesis will not be rejected if any of the CRs contains the origin. Based on the graphical summary in Figure 4, we conclude that the following six predictors are prognostic of overall survival: PAIN, ECOG, LDH.High, HGB, PSA, and ALKPHOS. The selected factors were also identified in other independent studies (Halabi et al., 2003; Scher et al., 2012; Ryan et al., 2013; Halabi et al., 2013), confirming the clinical importance of these factors on the patients' overall survival time.

5 Discussion

In this article, we propose a minimand perturbation method to derive the confidence region and estimate the covariance matrix for the regularized proportional hazards model. The theoretical background of applying the perturbation method to time-to-event outcomes is also justified in this article. Through simulations using the ALASSO penalty, we demonstrate that the proposed method successfully addresses many limitations of the existing variance estimator and provides a better standard error estimate than the sandwich formula. The simulation results also suggest the potential of applying the methodology to the adaptive elastic-net penalty. In a finite sample situation, similar to Zhang and Lu (2007)

and Minnier et al. (2011), we observe that the sandwich formula tends to under-estimate the true standard error of parameter estimate, which may lead to a liberal statistical inference. Our work complement the asymptotic results established by Zhang and Lu (2007) in the regularized Cox's model. Although the asymptotic properties are crucial for evaluating the merits of an estimation procedure, the finite sample inference has far-reaching impact on real practices. As a result, we propose novel improvements to alleviate the limitations of the current inference procedure.

The perturbation method requires essentially the same computation effort as that of the ALASSO-Cox. The computation and the demonstration codes are available via these two links: www.duke.edu/halab001/CommunicationsinStatistics/Demonstration.R www.duke.edu/halab001/CommunicationsinStatistics/pAlASSO.R In the first simulation example, it takes 1.4 and 3.5 seconds to run each perturbation using sample sizes of 200 and 800, respectively, and 20 predictors on a laptop PC with an Intel i5-2520M CPU and 4.0 GB memory. The number of perturbations is another pivotal factor in the computation. In our experience, the results varied slightly after $B = 300$ perturbations. To facilitate the computation speed, we recommend implementing parallel computing techniques (Tierney, Rossini, Li, & Sevcikova, 2012; Urbanek, 2011) to fully exploit the modern multi-thread computer hardware.

The scope of statistical inference ranges from point estimation to hypothesis testing. The confidence region complements the insufficiency of the point estimate and provides a thorough statistical inference. In medical research and in other disciplines, the point estimate is the first step of summarizing the association between the predictors and the outcome. Subsequent analysis, such as survival prediction and risk assessment, depends on both the point and interval estimates. Hence, it is vital to provide a comprehensive inference procedure that quantifies the uncertainty in the point estimate for applied practitioners.

The Cox's model will continue to be employed for developing prognostic models of time-to-event outcomes in clinical studies. To better promote the penalized regression methods for clinical applications, it is essential to disclose all the information relevant to clinical decision making. While a substantial amount of research has been dedicated to the theoretical and computational aspects of the penalized regression methods, there are numerous inferential questions that remain to be answered. As a result, the statistical inferences from the interval estimator is an evolving area of research.

In summary, we have presented a perturbation method which obtains the confidence region for the regularized Cox's model. The results of our simulations are favorable and demonstrate great finite sample performance across different scenarios. Moreover, this approach is easy to implement and can be extended to other penalty functions.

References

Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP-dimensionality. Annals of Statistics. 2011; 39:3092–3120. DOI: 10.1214/11-AOS911 [PubMed: 23066171]

Chatterjee A, Lahiri SN. Asymptotic properties of the residual bootstrap for Lasso estimators. Proceedings of the American Mathematical Society. 2010; 138:4497–4509. DOI: 10.1090/S0002-9939-2010-10474-4

Chatterjee A, Lahiri SN. Bootstrapping Lasso estimators. Journal of the American Statistical Association. 2011; 106:608–625. DOI: 10.1198/jasa.2011.tm10159

Cox DR. Regression models and life tables (with discussioon). Journal of the Royal Statistical Society, Ser B. 1972; 34:187–220.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with discussion). Annals of Statistics. 2004; 32:409–499.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360. DOI: 10.1198/016214501753382273

Fan J, Li R. Variable seleciton for Cox's proportional hazards model and frailty model. Annals of Statistics. 2002; 30:74–99. DOI: 10.1214/aos/1015362185

Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1:302–332. DOI: 10.1214/07-AOAS131

Friedman, J., Hastie, T., Tibshirani, R. Lasso and elastic-net regularized generalized linear models [Computer software manual]. 2012. Retrieved from http://CRAN.R-project.org/package=glmnet (R package version 1.8-2)

Halabi S, Lin CY, Small EJ, Armstrong AJ, Kaplan EB, Petrylak D, Sartor O. Prognostic model predicting metastatic castration-resistant prostate cancer survival in men treated with second-line chemotherapy. Journal of the National Cancer Institute. 2013; 105:1729–1737. DOI: 10.1093/jnci/djt280 [PubMed: 24136890]

Halabi S, Small EJ, Kantoff P, Kaplan EB, Dawson N, Levine E, Vogelzang N. Prognostic model for prediting survival in men with hormone-refractory metastatic prostate cancer. Journal of Clinical Oncology. 2003; 21:1232–1237. DOI: 10.1200/jco.2003.06.100 [PubMed: 12663709]

Ibrahim JG, Chen MM, MacEachern SN. Bayesian variable selection for proportional hazards models. Canadian Journal of Statistics. 1999; 27:701–717. DOI: 10.2307/3316126

Jin Z, Ying Z, Wei L. A simple resampling method by perturbing the minimand. Biometrika. 2001; 88:381–390. DOI: 10.1093/biomet/88.2.381

Kelly W, Halabi S, Carducci M, George D, Mahoney J, Stadler W, Small E. Randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer: CALGB 90401. Journal of Clinical Oncology. 2012; 30:1232–1237. DOI: 10.1200/jco.2011.39.4767

Knight K, Fu W. Asymptotics for Lasso-type estimators. Annals of Statistics. 2000; 28:1356–1378. DOI: 10.1214/aos/1015957397

Kosorok, M. Introduction to empirical process and semiparametric inference. Springer-Verlag; New York: 2008.

Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates. Journal of the American Statistical Association. 2011; 106:1371–1382. DOI: 10.1198/jasa.2011.tm10382 [PubMed: 22844171]

Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society, Ser B. 2007; 69:659–677. DOI: 10.1111/j.1467-9868.2007.00607.x

Ryan CJ, Smith MR, de Bono JS, Molina A, Logothetis CJ, de Souza P, Rathkopf DE. Abiraterone in metastatic prostate cancer without previous chemotherapy. New England Journal of Medicine. 2013; 368:138–148. DOI: 10.1056/NEJ-Moa1209096 [PubMed: 23228172]

Scher HI, Fizazi K, Saad F, Taplin ME, Sternberg CN, Miller K, de Bono JS. Increased survival with enzalutamide in prostate cancer after chemotherapy. 2012; 367:1187–1197. DOI: 10.1056/NEJMoa1207506

Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011; 39:1–13.

Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Ser B. 1996; 58:267–288.

Tibshirani R. The Lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16:385–395. [PubMed: 9044528]

Tierney, L., Rossini, AJ., Li, N., Sevcikova, H. Support for simple parallel computing in R [Computer software manual]. 2012. Retrieved from http://CRAN.R-project.org/package=snow (R package version 0.3-10)

Tsiatis AA. A large sample study of Cox's regression model. Annals of Statistics. 1981; 9:93–108. DOI: 10.1214/aos/1176345335

Urbanek, S. Parallel processing of R code on machines with multiple cores or CPUs [Computer software manual]. 2011. Retrieved from http://CRAN.R-project.org/package=multicore (R package version 0.1-7)

Wu Y. Elastic net for Cox's proportional hazards model with a solution path algorithm. Statistica Sinica. 2012; 22:271–294. DOI: 10.5705/ss.2010.107

Zhang HH, Lu W. Adaptive Lasso for Cox's proportioanl hazards model. Biometrika. 2007; 94:691–703. DOI: 10.1093/biomet/asm037

Zhao S, Li Y. Principled sure independence screening for cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis. 2012; 105:397–411. DOI: 10.1016/j.jmva.2011.08.002 [PubMed: 22408278]

Zou H. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429. DOI: 10.1198/016214506000000735

Zou H, Zhang HH. On the adaptive Elastic-Net with a divergingnumber of parameters. Annals of Statistics. 2009; 37:1733–1751. DOI: 10.1214/08-AOS625 [PubMed: 20445770]

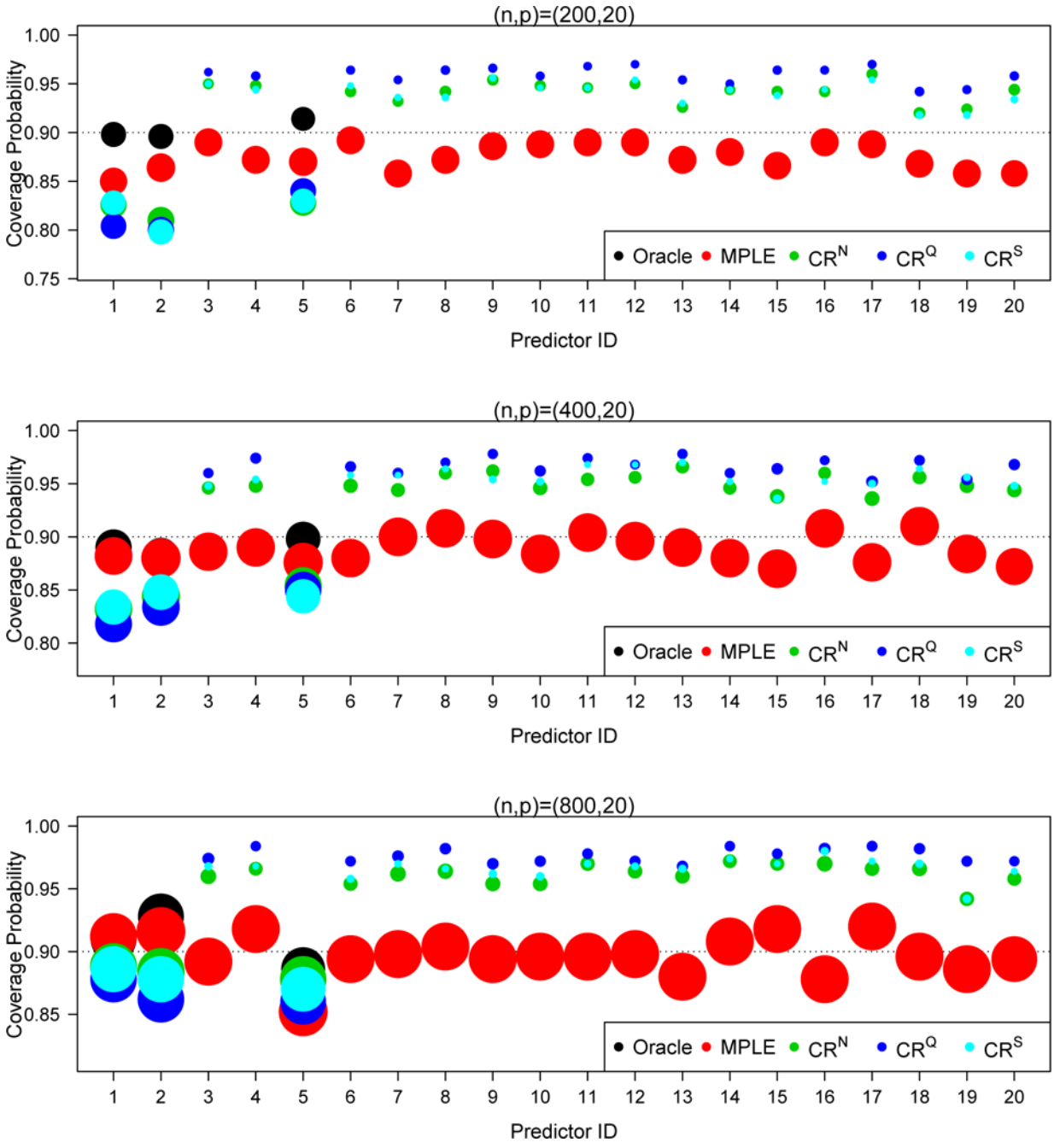

Figure 1.

Empirical coverage probabilities for Example 1 with 20 predictors. The diameters of the circles are proportional to the length of the confidence regions. The online version of this plot is in color.

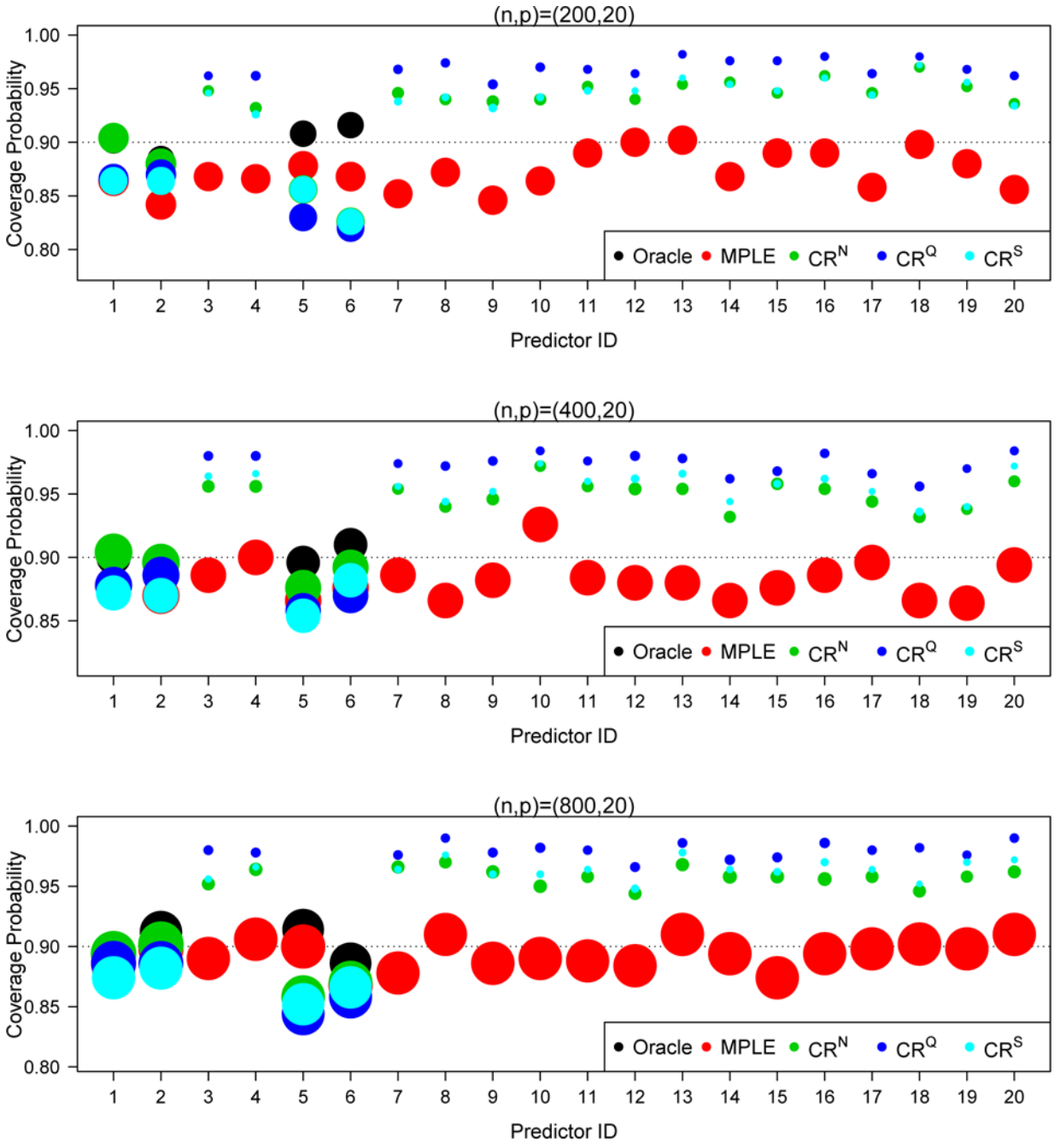

Figure 2.

Empirical coverage probabilities for Example 2 with 20 predictors. The diameters of the circles are proportional to the length of the confidence regions. The online version of this plot is in color.

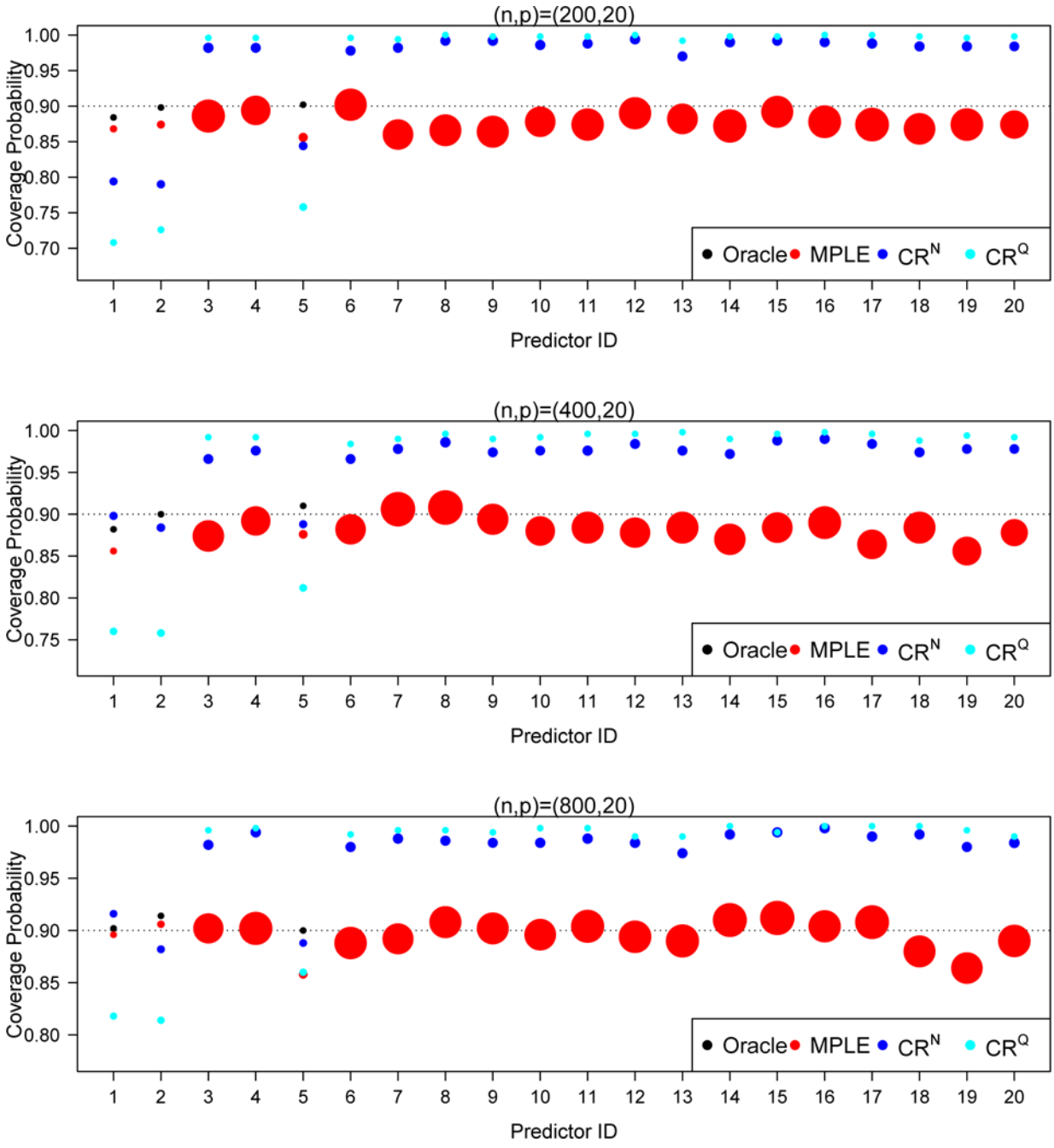

Figure 3.

Empirical coverage probabilities for Example 3 with 20 predictors. The diameters of the circles are proportional to the length of the confidence regions. The online version of this plot is in color.

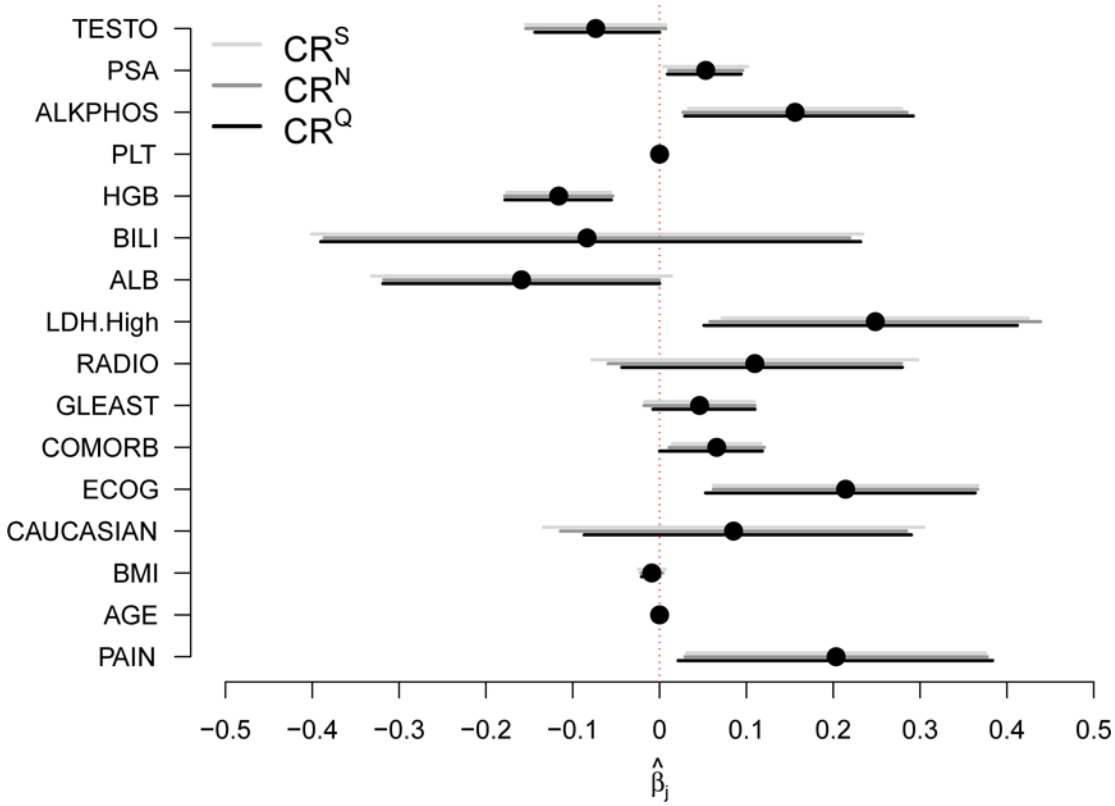

Figure 4.
90% CRs for a subset of CALGB 90401 data. The adaptive LASSO point estimates are marked with solid circles.

Table 1

Empirical coverage probabilities and length of confidence regions for Example 1 with 10 predictors.

	Oracle		MPLE		CR^S		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
					(n, p) = (200, 10)					
1	89	1.058	89	1.095	85	1.019	85	1.084	84	1.063
2	90	1.059	90	1.141	87	1.052	87	1.093	86	1.076
3	–	–	87	1.123	92	0.306	91	0.552	94	0.453
4	–	–	90	1.124	94	0.257	94	0.522	95	0.413
5	90	1.012	90	1.140	86	1.020	87	1.083	86	1.066
6	–	–	89	1.117	92	0.281	92	0.538	94	0.434
7	–	–	88	1.119	93	0.297	93	0.530	95	0.429
8	–	–	88	1.119	92	0.292	93	0.518	95	0.422
9	–	–	90	1.118	93	0.276	93	0.530	95	0.425
10	–	–	90	1.068	95	0.286	95	0.510	96	0.411
					(n, p) = (400, 10)					
1	91	0.740	91	0.753	88	0.743	88	0.757	87	0.746
2	90	0.739	89	0.783	86	0.748	87	0.765	86	0.753
3	–	–	90	0.767	95	0.160	95	0.328	97	0.252
4	–	–	88	0.767	95	0.183	94	0.334	96	0.261
5	90	0.706	87	0.783	87	0.723	88	0.748	87	0.737
6	–	–	89	0.767	94	0.170	94	0.342	96	0.267
7	–	–	90	0.769	95	0.160	95	0.317	95	0.249
8	–	–	91	0.767	95	0.150	94	0.325	96	0.248
9	–	–	91	0.770	96	0.141	95	0.315	97	0.239
10	–	–	88	0.736	95	0.160	94	0.302	96	0.235
					(n, p) = (800, 10)					
1	90	0.519	91	0.523	88	0.520	89	0.528	89	0.521
2	88	0.519	88	0.546	87	0.524	88	0.533	87	0.526
3	–	–	90	0.535	95	0.096	95	0.205	97	0.156

	Oracle		MPLE		CR^S		CR^N		CR^Q	
Predictor ID	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
4	–	–	89	0.535	94	0.106	93	0.209	96	0.160
5	92	0.495	91	0.544	88	0.505	90	0.520	89	0.512
6	–	–	88	0.535	97	0.101	95	0.205	98	0.158
7	–	–	89	0.535	95	0.103	94	0.209	96	0.160
8	–	–	89	0.535	95	0.091	94	0.199	96	0.151
9	–	–	90	0.535	96	0.090	95	0.198	97	0.149
10	–	–	91	0.512	97	0.084	96	0.189	98	0.141

CR^S: Confidence Region constructed by the sandwich formula.

CR^N: Confidence Region constructed by the normal approximation.

CR^Q: Confidence Region constructed by the empirical quantiles.

Table 2

Estimated and actual standard errors for estimates for Example 1.

	SE	Perturbation ASE	Sandwich ASE	SE	Perturbation ASE	Sandwich ASE
		(n, p) = (200, 10)			(n, p) = (200, 20)	
$\hat{\beta}_1$	0.379	0.329 (0.050)	0.310 (0.069)	0.402	0.336 (0.060)	0.315 (0.069)
$\hat{\beta}_2$	0.362	0.332 (0.049)	0.320 (0.056)	0.399	0.344 (0.057)	0.324 (0.059)
$\hat{\beta}_5$	0.363	0.329 (0.047)	0.310 (0.053)	0.396	0.335 (0.051)	0.314 (0.056)
		(n, p) = (400, 10)				
$\hat{\beta}_1$	0.232	0.230 (0.018)	0.226 (0.009)	0.267	0.240 (0.022)	0.226 (0.014)
$\hat{\beta}_2$	0.244	0.233 (0.019)	0.227 (0.009)	0.273	0.245 (0.020)	0.229 (0.014)
$\hat{\beta}_5$	0.244	0.227 (0.020)	0.220 (0.015)	0.261	0.238 (0.020)	0.221 (0.010)
		(n, p) = (800, 10)				
$\hat{\beta}_1$	0.160	0.161 (0.011)	0.158 (0.004)	0.161	0.162 (0.011)	0.158 (0.004)
$\hat{\beta}_2$	0.172	0.162 (0.011)	0.159 (0.005)	0.164	0.162 (0.011)	0.159 (0.005)
$\hat{\beta}_5$	0.156	0.158 (0.011)	0.154 (0.006)	0.164	0.159 (0.011)	0.154 (0.006)

SE: Sample standard error

ASE: Average of estimated standard error.

Table 3

Sure Screening Rate and Sure Screening Size of Example 1a.

(*n*, *p*)	Sure Screening Rate	Sure Screening Size
(200, 1000)	15.2	225.33 (229.21)
(400, 1000)	78.6	55.32 (99.02)
(800, 1000)	99.4	7.29 (15.47)

The numbers in the parentheses are standard deviation of sure screening size.

Table 4

Empirical coverage probabilities and length of confidence regions for Example 1a.

	Oracle		MPLE		CR^S		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
					(n, p) = (200, 10)					
1	88.4	1.061	80.1	1.248	77.8	1.091	80.1	1.266	79.3	1.228
2	89.2	1.058	89.9	1.299	79.7	1.123	80.6	1.285	83.2	1.244
5	99.2	1.007	82.8	1.236	77.2	1.105	81.5	1.269	81.5	1.229
					(n, p) = (400, 10)					
1	90.8	0.737	82.8	0.883	80.1	0.832	85.5	0.939	83.7	0.924
2	93.0	0.737	83.9	0.894	83.7	0.835	88.1	0.946	86.8	0.929
5	88.8	0.706	81.2	0.866	78.9	0.805	83.7	0.910	82.5	0.894
					(n, p) = (800, 10)					
1	91.8	0.519	83.9	0.607	84.9	0.579	87.3	0.637	88.2	0.630
2	91.2	0.519	84.5	0.613	83.0	0.583	87.0	0.643	85.6	0.634
5	89.8	0.495	78.0	0.603	77.2	0.568	81.4	0.623	82.2	0.614

CR^S: Confidence Region constructed by the sandwich formula.

CR^N: Confidence Region constructed by the normal approximation.

CR^Q: Confidence Region constructed by the empirical quantiles.

Table 5

Empirical coverage probabilities and length of confidence regions for Example 2 with 10 predictors.

	Oracle		MPLE		CR^S		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
					(n, p) = (200, 10)					
1	87.0	0.381	88.4	0.402	86.2	0.387	88.0	0.416	87.8	0.410
2	86.2	0.382	83.4	0.404	83.6	0.389	86.0	0.417	84.0	0.411
3	–	–	91.6	0.386	94.6	0.099	94.0	0.177	96.8	0.139
4	–	–	88.4	0.389	93.2	0.114	93.0	0.186	94.6	0.151
5	91.8	0.373	89.6	0.394	86.8	0.367	86.0	0.387	85.2	0.379
6	89.8	0.375	89.4	0.396	88.4	0.373	88.6	0.394	87.0	0.386
7	–	–	91.6	0.386	96.8	0.092	95.4	0.180	97.0	0.141
8	–	–	87.6	0.386	92.2	0.104	91.2	0.184	96.0	0.147
9	–	–	90.2	0.388	95.0	0.103	94.6	0.173	97.0	0.137
10	–	–	88.0	0.387	93.8	0.110	93.0	0.183	95.8	0.147
					(n, p) = (400, 10)					
1	88.4	0.264	89.2	0.275	88.2	0.267	89.0	0.281	88.2	0.277
2	88.2	0.265	88.8	0.277	86.4	0.268	88.8	0.284	87.6	0.279
3	–	–	88.0	0.265	95.8	0.067	94.4	0.115	98.2	0.090
4	–	–	89.6	0.264	95.8	0.061	94.6	0.107	97.0	0.082
5	90.0	0.260	89.8	0.271	87.0	0.263	88.6	0.271	86.8	0.266
6	89.0	0.260	88.2	0.272	86.8	0.263	88.8	0.272	86.8	0.266
7	–	–	90.6	0.266	94.4	0.062	93.4	0.113	97.6	0.088
8	–	–	88.4	0.266	95.0	0.054	93.4	0.109	97.2	0.085
9	–	–	90.2	0.264	94.8	0.053	94.2	0.102	97.2	0.077
10	–	–	89.4	0.266	94.2	0.060	92.8	0.117	96.4	0.092
					(n, p) = (800, 10)					
1	88.2	0.186	87.0	0.193	86.0	0.188	88.8	0.195	88.4	0.192
2	90.2	0.186	89.8	0.192	87.8	0.187	89.4	0.194	87.8	0.191
3	–	–	90.2	0.185	96.2	0.032	95.6	0.063	98.6	0.047

	Oracle		MPLE		CR^S		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
4	–	–	89.4	0.186	96.2	0.034	93.6	0.070	99.2	0.054
5	90.0	0.183	89.4	0.190	86.8	0.184	87.2	0.187	85.8	0.184
6	87.8	0.183	86.6	0.189	84.4	0.184	84.4	0.187	82.6	0.184
7	–	–	90.4	0.185	95.6	0.031	95.0	0.062	97.4	0.047
8	–	–	88.8	0.186	95.8	0.032	94.8	0.063	97.8	0.048
9	–	–	90.8	0.185	97.0	0.026	96.2	0.060	98.0	0.045
10	–	–	89.2	0.185	96.6	0.034	95.2	0.064	98.0	0.049

CR^S: Confidence Region constructed by the sandwich formula.

CR^N: Confidence Region constructed by the normal approximation.

CR^Q: Confidence Region constructed by the empirical quantiles.

Table 6

Empirical coverage probabilities and length of confidence regions for Example 3 with 10 predictors.

	Oracle		MPLE		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
				(n, p) = (200, 10)				
1	89.2	1.090	89.2	1.128	85.2	1.357	80.4	1.205
2	91.4	1.096	90.2	1.254	84.4	1.405	78.2	1.227
3	–	–	87.6	1.237	96.6	0.592	99.4	0.395
4	–	–	88.8	1.238	96.6	0.537	100.0	0.350
5	90.8	0.958	89.8	1.253	89.4	1.249	84.0	1.155
6	–	–	90.4	1.229	97.4	0.569	99.4	0.379
7	–	–	85.4	1.234	98.0	0.577	100.0	0.396
8	–	–	87.8	1.231	98.2	0.554	99.0	0.367
9	–	–	88.2	1.233	97.8	0.546	99.4	0.362
10	–	–	91.2	1.100	98.8	0.487	100.0	0.319
				(n, p) = (400, 10)				
1	91.0	0.761	89.6	0.774	91.0	0.909	80.8	0.861
2	88.8	0.765	88.8	0.861	89.4	0.942	81.6	0.885
3	–	–	90.2	0.845	97.4	0.384	98.0	0.259
4	–	–	89.4	0.846	96.6	0.398	98.4	0.270
5	89.2	0.667	89.0	0.861	88.6	0.819	85.6	0.786
6	–	–	91.4	0.845	96.6	0.391	98.6	0.254
7	–	–	89.6	0.849	97.2	0.372	98.8	0.245
8	–	–	89.2	0.846	96.8	0.375	98.8	0.245
9	–	–	90.0	0.850	98.0	0.379	99.4	0.253
10	–	–	89.4	0.758	96.6	0.354	98.8	0.234
				(n, p) = (800, 10)				
1	90.4	0.534	90.2	0.539	90.0	0.597	83.4	0.572
2	88.2	0.538	89.2	0.601	89.8	0.626	85.4	0.595
3	–	–	90.6	0.590	97.2	0.268	99.2	0.177

	Oracle		MPLE		CR^N		CR^Q	
Predictor ID	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**	**Coverage**	**Length**
4	–	–	88.4	0.590	96.4	0.281	99.6	0.183
5	90.4	0.469	89.2	0.599	90.0	0.551	88.2	0.534
6	–	–	90.4	0.590	97.0	0.268	98.8	0.173
7	–	–	89.6	0.591	96.8	0.272	99.4	0.180
8	–	–	88.2	0.590	97.8	0.271	98.4	0.180
9	–	–	89.4	0.590	97.8	0.265	99.6	0.173
10	–	–	91.4	0.528	99.0	0.234	99.6	0.149

CR^N: Confidence Region constructed by the normal approximation.

CR^Q: Confidence Region constructed by the empirical quantiles.

Table 7

Variable Description and Summary Statistics.

Variable	Description	Mean	SD
PAIN	Binary: 1 if the patient has baseline pain.	0.39	–
CAUCASIAN	Binary: 1 if the patient is Caucasian.	0.86	–
LDH.High	Binary: 1 if the patient's Lactate dehydrogenase is greater than normal value	0.36	–
RADIO	Binary: 1 if the patient has prior radiation therapy.	0.79	–
ECOG	Eastern Cooperative Oncology Group performance status of 0–2.	0.59/0.38/0.03[1]	–
COMORB	Number of Comorbidity	1.54	1.53
GLEAST	Gleason Score	7.61	1.32
ALKPHOS	Alkaline Phosphatase (U/L). In log scale.	4.92	0.73
ALB	Albumin (g/dL).	3.94	0.45
BILI	Bilirubin (mg/dL)	0.52	0.27
HGB	Hemoglobin (g/DL).	12.68	1.54
PLT	Peripheral Platelet Count ($\times 10^3$/uL)	258.22	78.28
PSA	Baseline Prostate Speci c Antigen (ng/mL). In log scale.	4.32	1.57
TESTO	Testosterone (ng/dL). In log scale.	2.75	0.98
BMI	Body Mass Index (kg/m^2).	29.72	5.16
AGE	Age (years).	68.64	8.81

[1]These three values represent the proportion of patients having performance status of 0, 1 and 2.