Title

A surrogate ℓ0 sparse Cox's regression with applications to sparse high-dimensional massive sample size time-to-event data

Journal

Statistics in Medicine, 39(6)

ISSN

0277-6715

Authors

Kawaguchi, Eric S
Suchard, Marc A
Liu, Zhenqiu
et al.

Publication Date

2020-03-15

DOI

10.1002/sim.8438

Peer reviewed

Published in final edited form as:

Stat Med. 2020 March 15; 39(6): 675–686. doi:10.1002/sim.8438.

A surrogate ℓ_0 sparse Cox's regression with applications to sparse high-dimensional massive sample size time-to-event data

Eric S. Kawaguchi[1], **Marc A. Suchard**[1,2,3], **Zhenqiu Liu**[4], **Gang Li**[1,2]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles, California

[2]Department of Biomathematics, University of California, Los Angeles, California

[3]Department of Human Genetics, University of California, Los Angeles, California

[4]Department of Public Health Sciences, Penn State Cancer Institute, Hershey, Pennsylvania

Abstract

Sparse high-dimensional massive sample size (sHDMSS) time-to-event data present multiple challenges to quantitative researchers as most current sparse survival regression methods and software will grind to a halt and become practically inoperable. This paper develops a scalable ℓ_0-based sparse Cox regression tool for right-censored time-to-event data that easily takes advantage of existing high performance implementation of ℓ_2-penalized regression method for sHDMSS time-to-event data. Specifically, we extend the ℓ_0-based broken adaptive ridge (BAR) methodology to the Cox model, which involves repeatedly performing reweighted ℓ_2-penalized regression. We rigorously show that the resulting estimator for the Cox model is selection consistent, oracle for parameter estimation, and has a grouping property for highly correlated covariates. Furthermore, we implement our BAR method in an R package for sHDMSS time-to-event data by leveraging existing efficient algorithms for massive ℓ_2-penalized Cox regression. We evaluate the BAR Cox regression method by extensive simulations and illustrate its application on an sHDMSS time-to-event data from the National Trauma Data Bank with hundreds of thousands of observations and tens of thousands sparsely represented covariates.

Keywords

Censoring; high-dimensional covariates; massive sample size; penalized regression; proportional hazards; survival analysis

---

Correspondence: Gang Li, Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA 90095. vli@ucla.edu.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Regularity conditions, proofs of theorems and additional simulation results are publicly available online in the supporting information tab of this article and implementation of BAR for right-censored time-to-event data can be found on the Github page [https://github.com/OHDSI/BrokenAdaptiveRidge].

1 | INTRODUCTION

Advancements in medical informatics tools and high-throughput biological experimentation are making large-scale data routinely accessible to researchers, administrators, and policymakers. This data deluge poses new challenges and critical barriers for quantitative researchers as existing statistical methods and software grind to a halt when analyzing these large-scale data sets, and calls for appropriate methods that can readily fit large-scale data. This paper primarily concerns survival analysis of sparse high-dimensional massive sample size (sHDMSS) data, a particular type of large-scale data with the following characteristics: (1) high-dimensional with a large number of covariates (p_n in thousands or tens of thousands), (2) massive in sample-size (n in thousands to hundreds of millions), (3) sparse in covariates with only a very small portion of covariates being nonzero for each subject, and (4) rare in event rate. An example of sHDMSS data is the pediatric trauma mortality data from the National Trauma Data Bank (NTDB) maintained by the American College of Surgeons.[1] This data set includes 210 555 patient records of injured children under 15 collected over 5 years from 2006 to 2010. Each patient record includes 125 952 binary covariates that indicate the presence or absence of an attribute (ICD9 Codes, AIS codes, etc) as well as their two-way interactions. The data matrix is extremely sparse with less than 1% of the covariates being non zero. The event (mortality) rate is also very low at 2%. Another application domain where sHDMSS data are common is drug safety studies that use massive patient-level databases such as the US FDA's Sentinel Initiative (https://www.fda.gov/safety/fdassentinelinitiative/ucm2007250.htm)and the Observational Health Data Sciences and Informatics program (https://ohdsi.org/) to study rare adverse events with hundreds of millions of patient records and tens of thousands of patient attributes that are sparse in the covariates.

The sHDMSS survival data present multiple challenges to quantitative researchers. First, not all of the thousands of covariates are expected to be relevant to an outcome of interest. It would also be practically undesirable to predict a patient outcome using thousands of covariates. Traditionally, researchers hand-pick subject characteristics to include in an analysis. However, hand picking can introduce not only bias, but also a source of variability between researchers and studies. Moreover, it would become impractical in large-scale evidence generation when hundreds or thousands of analyses are to be performed.[2] Hence, automated sparse regression methods are desired. Secondly, the commonly used "divide and conquer" strategy for massive size data is deemed inappropriate for sHDMSS time-to-event data since each of the divided data would have too few events for a meaningful analysis. Third, sHDMSS data presents a critical barrier to the application of existing sparse survival regression methods, since most current methods and standard software become inoperable for large data sets due to high computational costs and large memory requirements. Although many sparse survival regression methods are available,[3–10] to the best of our knowledge, only LASSO, Elastic Net[11] and ridge regression have been adapted to fit sHDMSS time-to-event data. In particular, Mittal et al[12] developed a tool, named CYCLOPS, for fitting LASSO and ridge Cox regression with sHDMSS time-to-event data by storing data in a sparse format, exploiting sparsity in the data and partial likelihood, and using multicore threading and vector processing, along with other high-performance computing techniques, which delivers > 10-fold speedup[12] over its competitors. However, ridge Cox regression does not yield a sparse model and LASSO tends to select too many noise features and is biased for estimation.[13,14] Improved sparse Cox regression tools for sHDMSS time-to-event data are desired.

The purpose of this paper is to develop a surrogate ℓ_0-based sparse Cox regression method and adapt it to sHDMSS time-to-event data. It is well known that ℓ_0-penalized regression is natural for variable selection and parameter estimation with some optimal properties.[15–18] On the other hand, it is also known to have some pitfalls such as instability[19] and being unscalable to even moderate dimensional covariates. The broken adaptive ridge (BAR) estimator, defined as the limit of an iteratively reweighted ℓ_2-penalization algorithm, was introduced to approximate the ℓ_0-penalization problem and has been recently shown to possess some desirable selection, estimation, and clustering properties under the linear model and several other model settings.[10,20–22] It is also computationally scalable to high-dimensional covariates and stable for variable selection as discussed later in Remark 2 of Section 2. However, the BAR method has yet to be rigorously studied for the Cox model. Moreover, current BAR algorithms have only been implemented for densely-represented covariates and are unsuitable for sHDMSS data due to high computational costs, high memory requirements, and numerical instability. Computation of the Cox partial likelihood and its derivatives is particularly demanding for massive sample size data since the required number of operations grows at the rate of $O(n^2)$. The key contributions of this paper are twofold. First, we rigorously extend the BAR methodology to the Cox model. Specifically, we establish the selection consistency, an oracle property for parameter estimation, and a grouping property of highly correlated covariates for the Cox model. It is worth noting that the theoretical extension of the BAR methodology to Cox model is nontrivial and notably different from other models because the log-partial likelihood for the Cox model is not the sum of independent terms and the standard martingale central limit theorem used to derive the asymptotic theory for Cox's model with a fixed number of covariates is no longer applicable when the number of parameters diverges. Furthermore, because BAR involves performing an infinite number of penalized regressions, the derivations of its selection consistency and oracle property for estimation are substantially different from those for a single-step oracle estimator in the literature. The second key contribution of this paper is to develop an efficient implementation of BAR for Cox regression with sHDMSS time-to-event data by leveraging existing efficient massive ℓ_2-penalized Cox regression techniques,[12] which include employing a column relaxation with logistic loss (CLG) algorithm using one-dimensional updates and a one-step Newton-Raphson approximation as well as exploiting the sparsity in the covariate structure and the Cox partial likelihood to reduce the number of operations from $O(n^2)$ to $O(n)$.

In Section 2, we formally define the BAR estimator, state its theoretical properties for variable selection, parameter estimation, and grouping highly correlated covariates for the Cox model, and describe an efficient implementation for sHDMSS survival data. We also discuss how to adapt BAR as a postscreening sparse regression method for ultrahigh dimensional Cox regression with relatively small sample size. In Section 3, we present simulation studies to demonstrate the performance of the CoxBAR estimator with both moderate and massive sample size in various low and high-dimensional settings. We provide a real data example using the pediatric trauma mortality data[12] in Section 4. Lastly, we give closing remarks in Section 5. The appendix collects proofs of the theoretical results and regularity conditions needed for the derivations. An R package has been developed for BAR and made available at https://github.com/OHDSI/BrokenAdaptiveRidge.

2 | METHODOLOGY

2.1 | Cox's BAR regression and its large sample properties

2.1.1 | The data structure, model, and estimator—

Suppose that one observes a random sample of right-censored time-to-event data consisting of n independent and identically distributed triplets $\{(X_i, \delta_i, \mathbf{z}_i(\cdot))\}_{i=1}^n$, where for subject i, $X_i = \min(T_i, C_i)$ is the observed event time, $\delta_i = I(T_i \le C_i)$ is the censoring indicator, T_i is the event time of interest, and C_i is a censoring time that is conditionally independent of T_i given a p_n-dimensional, possibly time-dependent, covariate vector $\mathbf{Z}_i(\cdot) = \left(z_{i1}(\cdot), \ldots, z_{ip_n}(\cdot)\right)'$.

Assume the Cox[23] proportional hazard model

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

where $h\{t \mid \mathbf{z}(t)\}$ is the conditional hazard function of T_i given $\{\mathbf{z}(u), 0 \le u \le t, \}$, $h_0(t)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_{p_n}\right)$ is a vector of time-independent regression coefficients. Denote by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ the first q_n and remaining $p_n - q_n$ components of $\boldsymbol{\beta}$, respectively, and define $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ as the true values of $\boldsymbol{\beta}$, where, without loss of generality, $\boldsymbol{\beta}_{01} = \left(\beta_{01} \ldots, \beta_{0q_n}\right)$ is a vector of q_n nonzero values and $\boldsymbol{\beta}_{02} = \mathbf{0}$ is a $p_n - q_n$ dimensional vector of zeros. Further technical assumptions for $\boldsymbol{\beta}_0$ and p_n are given later in condition (C6) of Section S4 of the Supplementary Material. For simplicity, we work on the time interval $s \in [0, 1]$ as in the work of Andersen and Gill,[24] which can be extended to any time interval $[0, \tau]$ for $0 < \tau < \infty$. Using the standard counting process notation, the log-partial likelihood for the Cox model is defined as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^1 \boldsymbol{\beta}'\mathbf{z}_i(s)dN_i(s) - \int_0^1 \log\left[\sum_{j=1}^{n} Y_j(s)\exp\left\{\boldsymbol{\beta}'\mathbf{z}_j(s)\right\}\right] d\overline{N}(s), \tag{2}$$

where, for subject i, $Y_i(s) = I(X_i \ge s)$ is the at-risk process and $N_i(s) = I(X_i \le s, \delta_i = 1)$ is the counting process of the uncensored event with intensity process $h_i(t \mid \boldsymbol{\beta}) = h_0(t)\, Y_i(t)$ $\exp\{\mathbf{z}_i(t)'\boldsymbol{\beta}\}$ and $\overline{N} = \sum_{i=1}^n N_i$.

Our Cox's BAR estimation of $\boldsymbol{\beta}$ starts with an initial Cox ridge regression estimator[25]

$$\widehat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}}\left\{-2l_n(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p_n} \beta_j^2\right\}, \tag{3}$$

which is updated iteratively by a reweighed ℓ_2-penalized Cox regression estimator

$$\widehat{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\left(\widehat{\beta}_j^{(k-1)}\right)^2} \right\}, \quad k \geq 1, \tag{4}$$

where ξ_n and λ_n are nonnegative penalization tuning parameters. The BAR estimator is defined as

$$\widehat{\boldsymbol{\beta}} = \lim_{k \to \infty} \widehat{\boldsymbol{\beta}}^{(k)}. \tag{5}$$

Since ℓ_2-penalization yields a nonsparse solution, defining the BAR estimator as the limit is necessary to produce sparsity. Although λ_n is fixed at each iteration, it is weighted inversely by the square of the ridge regression estimates from the previous iteration. Consequently, coefficients whose true values are zero will have larger penalties in the next iteration, whereas penalties for truly nonzero coefficients will converge to a constant. We will show later in Theorem 1 that, under certain regularity conditions, the estimates of the truly zero coefficients shrink toward zero while the estimates of the truly nonzero coefficients converge to their oracle estimates with probability tending to 1. As illustrated by a small simulation in Section S2 (Figure S1) of the Supplementary Material, the signal (nonzero coefficients) and noise (zero coefficients) can be quickly separated within a few BAR iterations, although more iterations may be necessary in some scenarios to improve estimation of the nonzero coefficients.

***Remark* 1. (Computational aspects of BAR):** For moderate size data, one may calculate $\widehat{\boldsymbol{\beta}}^{(k)}$ in (4) using the Newton-Raphson method as in the work of Frommlet and Nuel,[26] who outlined an iterative reweighted ridge regression for generalized linear models. It appears at first sight that (4) will encounter numerical overflow as some of the coefficients $\widehat{\beta}_j^{(k-1)}$ will go to zero as k increases. However, it can be shown that after some simple algebraic manipulation the Newton-Raphson updating formula will only involve multiplications, instead of divisions, by $\widehat{\beta}_j^{(k-1)}$s, and thus numerical overflow can be avoided. Further details are provided in Section S1 (Equation (3)) of the Supplementary Material. We also note that because the limit of the BAR algorithm cannot be numerically achieved at any finite iteration step, an extra thresholding rule for small coefficients will be required to numerically obtain a sparse solution. However, this thresholding level can be set arbitrarily small (by default, we set the threshold value to 10^{-6} in our implementation) since it is simply used for numerical convergence to zero and has minimal impact on the resulting BAR estimator. Furthermore, Equation (3) of Section S1 of the Supplementary Material implies that, once a $\widehat{\beta}_j^{(k-1)}$ becomes zero, it will remain as zero in subsequent iterations. Thus, one only needs to update $\widehat{\boldsymbol{\beta}}^{(k)}$ within the reduced nonzero parameter space, an appealing computational advantage for high-dimensional settings.

For massive-size data with large n and p_n, the Newton-Raphson procedure, which at each iteration, calls for the calculation of both the gradient and Hessian can become practically infeasible due to high computational costs, memory requirements, and numerical instability. In Section 2.2, we will discuss how to adapt an efficient algorithm for massive ℓ_2-penalized Cox regression via cyclic coordinate descent and exploit the sparsity of the covariate structure to make BAR scalable to sHDMSS data.

***Remark* 2. (Broken adaptive ridge versus best subset selection):** The BAR method can be viewed as a performing a sequence of surrogate ℓ_0-penalizations, where each reweighted ℓ_2 penalty serves as a surrogate ℓ_0-penalty and the approximation of ℓ_0-penalization improves with each iteration. Consequently, BAR enjoys the best of ℓ_0- and ℓ_2-penalized regressions. For example, we establish in the next two sections that BAR possesses the oracle properties for estimation and selection consistency (an ℓ_0 property) as well as a grouping property (an ℓ_2 property). Numerically, for a fixed tuning parameter value, BAR is a surrogate to ℓ_0-penalization is not expected to be identical, but can be similar to the exact global ℓ_0-penalization solution where the latter must be solved using the best subset search (BSS). We illustrate this in Section S3 of the Supplementary Material (Figures S2 and S3) using a small simulation study. It is worth emphasizing that BAR overcomes some shortcomings of BSS; for example, BSS is computationally NP-hard and can be unstable for variable selection,[19] whereas BAR is scalable to high-dimensional covariates and is more stable for variable selection as demonstrated in Figures S2 and S3 in Section S3 of the Supplementary Material.

2.1.2 | Oracle properties—We establish the oracle properties for the BAR estimator for simultaneous variable selection and parameter estimation where we allow both q_n and p_n to diverge to infinity.

Theorem 1 (Oracle properties).: Assume the regularity conditions (C1) to (C6) in Section S4.1 of the Supplementary Material hold. Let $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ be the first q_n and the remaining $p_n - q_n$ components of the BAR estimator $\widehat{\boldsymbol{\beta}}$, respectively. Then, as $n \to \infty$, with probability tending to one,

- **a.** the BAR estimator $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right)$ exists and is unique, where $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$;
- **b.** $\sqrt{n}\mathbf{b}_n' \Sigma(\boldsymbol{\beta}_0)_{11}^{-1/2}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\right) \xrightarrow{D} N(0, 1)$, for any q_n-dimensional vector $\mathbf{b}_n$ such that $\|\mathbf{b}_n\|_2 \le 1$ and where $\Sigma(\boldsymbol{\beta}_0)_{11}$ is the first $q_n \times q_n$ submatrix of $\Sigma(\boldsymbol{\beta}_0)$, where $\Sigma(\boldsymbol{\beta}_0)$ is defined in Condition (C4).

Theorem 1(a) establishes selection consistency of the BAR estimator. Part (b) of the theorem essentially states that the nonzero component of the BAR estimator is asymptotically normal and equivalent to the weighted ridge estimator of the oracle model, as shown in the proof provided in Section S4.2 of the Supplementary Material.

***Remark* 3 (Ultrahigh-dimensional covariates setting).:** Although we allow p_n to diverge, the asymptotic properties of the BAR estimator in the Section 2.1 are derived for $p_n < n$. In an ultrahigh-dimensional setting where the number of covariates far exceeds the number of observations ($p_n \gg n$), one may couple a sure screening[27] method with the BAR estimator to obtain a two-step estimator with desirable selection and estimation properties. The orders of q_n, p_n, and n and their relationships depend on the employed screening procedure. For example, coupling the BAR estimator with the sure joint screening procedure[28] has been explored in the work of Kawaguchi.[29]

2.1.3 | A grouping property—

When the true model has a group structure, it is desirable for a variable selection method to either retain or drop all variables that are clustered within the same group. It is well known that ridge regression possesses the grouping property for highly correlated covariates.[11] Because the BAR estimator is based on an iterative ridge regression, we show that BAR also possesses a grouping property for highly correlated covariates as stated in following theorem.

Theorem 2.: Let λ_n, $\{(X_i, \delta_i, \mathbf{z}_i)\}_{i=1}^{n}$ be given and assume that $Z = (\mathbf{z}_i', \ldots \mathbf{z}_n')$ is standardized. That is, for all $j = 1, \ldots, p_n$, $\sum_{i=1}^{n} z_{ij} = 0, \mathbf{z}_{[,j]}'\mathbf{z}_{[,j]} = n - 1$, where $\mathbf{z}_{[,j]}$ is the j^{th} column of Z. Suppose the regularity conditions (C1) to (C6) in Section S4.1 of the Supplementary Material hold and let $\hat{\boldsymbol{\beta}}$ be the BAR estimator. Then, for any $\hat{\beta}_i \neq 0$ and $\hat{\beta}_j \neq 0$,

$$\left|\hat{\beta}_i^{-1} \pm \hat{\beta}_j^{-1}\right| \leq \frac{1}{\lambda_n}\sqrt{2\left\{(n-1)\left(1 \pm r_{ij}\right)\right\}}\sqrt{n\left(1 + d_n\right)^2}, \tag{6}$$

with probability tending to one, where $d_n = \sum_{i=1}^{n} \delta_i$, and $r_{ij} = \frac{1}{n-1}\mathbf{z}_{[,i]}'\mathbf{z}_{[,j]}$ is the sample correlation of $\mathbf{z}_{[,i]}$ and $\mathbf{z}_{[,j]}$.

The proof is provided in Section S4.3 of the Supplementary Material. It is seen from (6) that, as $r_{ij} \rightarrow 1$, the absolute difference between $\hat{\beta}_i$ and $\hat{\beta}_j$ approaches 0, implying that the estimated coefficients of two highly positively correlated variables will be similar in magnitude. Similarly, the estimated coefficients of two highly negatively correlated variables are also similar in magnitude with a sign change.

2.1.4 | Selection of tuning parameters—

Model complexity depends critically on the choice of the tuning parameters. The BAR estimator depends on two tuning parameters, ie, ξ_n for the initial ridge estimator in (3) and λ_n for the iterative ridge step in (4). Our simulations in Section 3.1 illustrate that, while fixing λ_n, the BAR estimator is insensitive to the choice of ξ_n over a wide interval (Figure 1) and thus practically only optimization with respect to λ_n is needed.

We optimize with respect to λ_n in a similar manner to currently used penalization methods. A popular strategy for tuning parameter selection is to perform optimization with respect to a data-driven selection criterion such as cross-validation (CV),[30,31] Akaike information criterion,[15] and Bayesian information criterion (BIC).[16,17,32] Although CV has been used extensively in the literature, it has been known to asymptotically overfit models with a positive probability.[33,34] Recent theoretical work has shown that, for penalized Cox models that possess the oracle property, BIC-based tuning parameter selection identifies the true model with probability tending to one.[32] Further discussion on selecting λ_n for BAR is provided in the last paragraph of Section 3.2.

2.2 | Implementation of BAR for sHDMSS data

As mentioned in Remark 1, the Newton-Raphson algorithm used for each iteration of the BAR algorithm will become infeasible in large-scale settings with large n and p_n due to high computational costs, high memory requirements, and numerical instability. Furthermore, recently proposed BAR algorithms, as with most popularly available procedures, cannot directly handle sHDMSS data due to the computational burden imposed by the design matrix. Because BAR only involves fitting a reweighted Cox's ridge regression at each iteration step, it allows us to adapt an efficient algorithm developed by Mittal et al[12] for massive Cox ridge regression.

2.2.1 | Adaptation of existing efficient algorithms for fitting massive ℓ_2-penalized Cox's regression—

Mittal et al[12] developed an efficient implementation of the massive Cox's ridge regression for sHDMSS data. For parameter estimation, the authors adopted the CLG algorithm of Zhang and Oles,[35] which is a type of cyclic coordinate descent algorithm that estimates the coefficients using one-dimensional updates. The CLG easily scales to high-dimensional data[7,36,37] and has been recently implemented for fitting ℓ_2- and ℓ_1-penalized generalized linear models,[38] parametric time-to-event models,[39] and Cox's model.[12] Readers are encouraged to refer to Section S3 of the Supplementary Material for a detailed explanation of the algorithm.

The design matrix Z for sHDMSS data has few nonzero entries for each subject. Storing such a sparse matrix as a dense matrix is inefficient and may increase computation time and/or cause standard software to crash due to insufficient memory allocation. To the best of our knowledge, popular penalization packages such as **glmnet**[40] and **ncvreg**[41] do not support a sparse data format as an input for right-censored time-to-event models, although the former supports the input for other generalized linear models. For sHDMSS data, we propose to use specialized column-data structures as in the works of Mittal et al[12] and Suchard et al.[38] The advantage of this structure is two-fold, ie, it significantly reduces the memory requirement needed to store the covariate information, and performance is enhanced when employing cyclic coordinate descent. For example, when updating β_j, efficiency is gained when computing and storing the inner product $r_i = \mathbf{z}_i'\boldsymbol{\beta}$ using a low-rank update $r_i^{(new)} = r_i + z_{ij} + \Delta\beta_j$ for all i.[12,35,36,38,42]

Furthermore, one requires a series of cumulative sums introduced through the risk set $R_i = \{j : X_j > X_i\}$ for each subject i to calculate the gradient and Hessian diagonal. These cumulative sums would need to be calculated when updating each parameter estimate in the optimization routine. This can prove to be computationally costly, especially when both n and p_n are large. By taking advantage of the sparsity of the design matrix, one can reduce the computational time needed to calculate these cumulative sums by entering into this operation only if at least one observation in the risk set has a nonzero covariate value along dimension j and embarking on the scan at the first nonzero entry rather than from the beginning. Mittal et al[12] and Suchard et al[38] have implemented these efficiency techniques for conditional Poisson regression and Cox's regression, respectively. Our BAR implementation naturally exploits the sparsity in the design matrix and the partial likelihood by embedding an adaptive version of Mittal et al's[12] massive Cox's ridge regression within each iteration of the iteratively reweighted Cox's ridge regression.

3 | SIMULATIONS

This section presents three simulation studies. First, we demonstrate in Section 3.1 that, for fixed λ_n, the BAR estimator is insensitive to the tuning parameter ξ_n of its initial ridge estimator and does well in terms of performing variable selection and correcting possible bias of the initial ridge estimator. Then, in Section 3.2, we evaluate and compare the operating characteristics of BAR with some popular penalized Cox regression methods, where we only consider settings with moderate sample sizes due to the fact that most of the competing methods are inoperable for massive sample size data. Finally, in Section 3.3, we use a sHDMSS setting to illustrate the performance of BAR over its closest competitor.

Sections 3.1 and 3.2 employ the same simulation structure. Event times are drawn from an exponential proportional hazards model with baseline hazard $h_0(t) = 1$, $\beta_0 = (0.40, 0, 0.45, 0, 0.50, 0.55, 0, 0, 0.70, 0.80, \mathbf{0}_{p_n-10})$, representing $q_n = 6$ small to moderate effect sizes; the design matrix $Z = (\mathbf{z}_1', \dots, \mathbf{z}_n')$ is generated from a p_n-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$ with an autoregressive structure such that $\sigma_{ij} = 0.5^{|i-j|}$ and independent censoring times are generated from uniform distribution $U(0, u_{\max})$, where $u_{\max}$ is chosen to achieve different percentages of censoring. We describe how we simulate sHDMSS time-to-event data in Section 3.3.

3.1 | Broken adaptive ridge estimator for varying values of ξ_n

We illustrate how the BAR estimator behaves by fixing λ_n and varying the tuning parameter ξ_n of the initial Cox ridge regression in the following. Figures 1B to 1D depict the solution path plots average over 100 Monte Carlo simulations of the BAR estimator with respect to ξ_n over a wide interval $[10^{-2}, 10^{2}]$ for $n = 300$, $p_n = 100$, $\approx 25\%$ censoring, and $\lambda_n = \log(p_n)$, $0.5\log(p_n)$, $0.75\log(p_n)$, respectively. It is seen that the resulting BAR estimator is essentially unchanged, regardless of the choice of λ_n, over a large interval of ξ_n, suggesting that the BAR estimator is relatively insensitive to original ridge estimator.

As a reference, we also display the solution path plots of the corresponding initial ridge estimator in panel (a). The initial ridge estimator starts to introduce over shrinkage and, consequently, estimation bias when ξ_n exceeds 10. However, its bias has been effectively corrected by BAR. Therefore, by iteratively refitting reweighted Cox ridge regression, the BAR estimator not only performs variable selection by shrinking estimates of the true zero parameters to zero, but also effectively corrects the estimation bias from the initial Cox ridge estimator. Similar results are obtained for several different simulation scenarios and can be found in Section S4 of the Supplementary Material.

3.2 | Model selection and parameter estimation

In this simulation, we evaluate and compare the variable selection and parameter estimation performance of BAR with four popular penalized Cox regression methods, ie, LASSO,[3] SCAD,[4] adaptive LASSO (ALASSO),[5] and MCP.[6] We fix $\xi_n = 1$ for the BAR methods since

Section 3.1 yields evidence that the BAR estimator is insensitive to the selection of ξ_n. For all methods, a 25-value grid was used to find the optimal value of the tuning parameter via BIC minimization.[32]

Estimation bias is summarized through the mean squared bias, $E\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2\right)$. Variable selection performance is measured by a number of indices, ie, the mean number of false positives (FP), the mean number of false negatives (FN), and average similarity measure for support recovery where $SM = \|\hat{\mathcal{S}} \cap \mathcal{S}_0\|_0 / \sqrt{\|\hat{\mathcal{S}}|_0 \cdot \|\mathcal{S}_0\|_0}$ and $\mathcal{S}_0$ and $\hat{\mathcal{S}}$ are the set of indices for the nonzero components of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, respectively.[43] The similarity measure can be viewed as a continuous measure for true model recovery, ie, it is close to 1 when the estimated model is similar to the true model and close to 0 when the estimated model is highly dissimilar to the true model. We use the R package **ncvreg** to perform LASSO, ALASSO, SCAD, and MCP penalizations in our simulations. For ALASSO, we let the initial weight be the maximum partial likelihood estimator since $p_n < n$. Partial simulation results are summarized in Table 1 where we fix $n = 300, 1000$, $p_n = 100$, a censoring rate of ≈ 25%, and average results over 100 replications.

From Table 1, we have that, when the tuning parameter λ_n is selected by minimizing the BIC score as the other methods, the performance of BAR (BIC) is generally comparable to other methods with respect to all measures across both scenarios. We have conducted more extensive simulations with different combinations of model dimension, censoring rates, sample sizes, and model sparsity, which yielded consistent findings and are reported in Section S5 of the Supplementary Material.

Since BAR aims to approximate ℓ_0-penalized regression, it directly provides a surrogate optima to some popular information criteria with some prefixed λ_n. For example, performing BAR with $\lambda_n = c\log(p_n)$ for some $c > 0$ leads to a surrogate optima for the directly optimizing the extended BIC.[44–46] For thoroughness, in addition to using a 25-value grid for c, we also include simulation results in Table 1 for BAR with some prefixed values $\lambda_n = 0.5\log(p_n)$ and $\lambda_n = \log(p_n)$. Not surprisingly, BAR with these prefixed values produced sometimes slightly suboptimal, but generally comparable estimation and selection performance. We also conducted further simulations using a 10-value coarse grid for λ_n. The results are presented in Tables S1 to S3 of the Supplementary Material, which showed that the 10-value grid worked as well as the 25-value grid across almost all of our simulation scenarios. This suggests that potential computational savings could be gained for BAR by using either prefixed or a coarse grid of values for λ_n for massive data, which is also illustrated in Section 4 (Table 3).

3.3 | Sparse high-dimensional massive sample size data

In this simulation, we simulate a sHDMSS time-to-event data set with $n = 200,000$, $p_n = 20,000$, and $q_n = 80$. Event times are generated from an exponential hazards model with baseline hazard $h_0(t) = 1$, regression coefficients $\boldsymbol{\beta}_0 = (\mathbf{0.7}_{10}, \mathbf{0.5}_{10}, \mathbf{0.8}_{10}, \mathbf{1}_{10}, -\mathbf{0.7}_{10}, -\mathbf{0.5}_{10}, -\mathbf{0.8}_{10}, -\mathbf{1}_{10}, \mathbf{0}_{p_n-80})$, and a censoring rate of 95%. The covariates for each subject are simulated such, on average, 2% are assigned a nonzero value. The amount of memory used to store this dense design matrix would require over 16 GB, which exceeds the functional capacity of most statistical software packages on standard hardware. To overcome this difficulty, we efficiently store the information in a coordinate list fashion and compare our massive Cox's regression for BAR (mBAR) with the massive sparse Cox's regression for LASSO (mCox-LASSO) using the **Cyclops** package,[12,38] which, to the best of our knowledge, is the fastest software available today that exploits the sparsity of sHDMSS time-to-event data for efficient computing and offers > 10-fold speedup[12] over its competitors such as **CoxNet**[7] and **FastCox**.[47] For LASSO, CV (mCox-LASSO (CV)), combined with a nonconvex optimization technique which is more efficient than the classical grid search approach, and BIC score minimization (mCox-LASSO (BIC)), implemented with the classical grid search approach, were used to find the optimal value for the tuning parameter. For the mBAR method, we implement BIC score minimization using a grid search and two prefixed tuning parameters $\lambda_n = 0.5 \log(p_n)$ and $\log(p_n)$ for comparative purposes. We report the bias $\left(\|\hat{\beta} - \beta_0\|_2\right)$, number of FP, FN, and BIC score $\left(-2l_n(\hat{\boldsymbol{\beta}}) + \log(n)\sum_j I\left(\hat{\boldsymbol{\beta}}_j \neq 0\right)\right)$ in Table 2.

We observe that both mCox-LASSO methods have retained all 80 true nonzero coefficients together with a moderate to large number of noise variables (12 for BIC and 967 for CV). In contrast, mBAR (BIC) chooses a sparser model selecting all 80 nonzero coefficients and 5 noise variables. As expected, mBAR (BIC) is less biased (0.82) than mCox-LASSO (2.49 for BIC and 2.02 for CV) and has a much lower BIC score when compared to both mCox-LASSO methods. We also notice that mBAR with the two prefixed λ_n tends to underestimate the true model, ie, fixing $\lambda_n = \log(p_n)$ results in estimating a model that is too sparse, whereas $\lambda_n = 0.5 \log(p_n)$ produces a model that is closer to the oracle model.

We further examined the solution paths of mCox-LASSO and mBAR in Figure 2. The vertical solid and dashed lines in the mCox-LASSO solution path plot (Figure 2A) represent the estimates at the optimal tuning parameter obtained via CV and BIC minimization, respectively. We can see that the mCox-LASSO solution path changes rapidly as its tuning parameter varies and shows severe bias. In contrast, the mBAR solution path plot (Figure 2B) with respect to λ_n changes very slowly where the vertical line represents the estimates at the optimal tuning parameter selected by BIC minimization and selects a model with estimates that are less biased than mCox-LASSO (see Table 2). Furthermore, the optimal value of λ_n that minimizes the BIC score for mBAR roughly corresponds to 0.3 $\log(p_n)$. Since our empirical results suggest that the optimal value for λ_n generally lies within some constant of $\log(p_n)$, we recommend that a coarse grid search within $c \log(p_n)$ where $c \in (0, 1]$ can be used. This is further corroborated by additional simulations in the Supplementary Material (Tables S1 to S3).

For the mBAR method, we also made a solution path plot with respect to ξ_n, while fixing $\lambda_n = \log(p_n)$ in Figure 2C. It shows that the mBAR estimates are very stable over a large range of ξ_n, affirming our observation in Section 3.1 with small scale data that mBAR is generally insensitive ξ_n.

4 | PEDIATRIC TRAUMA MORTALITY

For an application of mBAR regression in the sHDMSS setting, we consider a subset of the NTDB, a trauma database maintained by the American College of Surgeons.[1] This data set was previously analyzed by Mittal et al[12] as an example for efficient massive Cox regression with mCox-LASSO and ridge regression to sHDMSS data. The data set includes 210 555 patient records of injured children under 15 that were collected over 5 years (2006 to 2010). Each patient record includes 125 952 binary covariates, which indicate the presence or absence of an attribute (ICD9 Codes, AIS codes, etc) as well as the two-way interactions. The outcome of interest is mortality after time of injury. The data is extremely sparse, with less than 1% of the covariates being nonzero and has a censoring rate of 98%. We randomly split the data into training and test sets of 168 000 and 42 555, respectively. The mortality rate of both sets were approximately equal to the combined rate. Similar to Section 3.3, we were unable to load the training set ($n = 168\ 000$, $p_n = 125\ 000$) into other popular oracle procedures due to the memory requirements needed to support a dense design matrix of that size and compare mBAR to mCox-LASSO. The BIC-score minimization over a penalization path of 10 tuning parameters was used to select the final model for both mBAR (fixing $\xi_n = \log(p_n)$) and mCox-LASSO. In addition, we perform mCox-LASSO using CV and mBAR with fixed tuning parameters $\lambda_n = 0.5\log(p_n)$ and $\log(p_n)$. The BIC score based on the training data is used to compare selection performance between models and discriminatory performance is measured using Harrell's c-statistic[48,49] based on the test data.

Table 3 summarizes the findings for our example, which reflect what we observe in Section 3.3. Massive Cox's regression for BAR, using BIC minimization, selects fewer covariates than both mCox-LASSO methods. Both model selection and discriminatory performance are similar to slightly superior for mBAR (BIC) over both mCox-LASSO methods. Again, mBAR with prefixed λ_n selects far fewer covariates than mBAR (BIC); however, the overall high c-index for both methods suggest that the strong predictors for pediatric trauma are still retained in the model. In terms of runtime, mBAR (BIC) is more time consuming than LASSO (BIC) as expected, but BAR with a prefixed tuning parameter value can help to reduce the runtime with a comparable prediction performance.

5 | DISCUSSION

We have extended the BAR methodology to Cox's model as a new sparse Cox regression method and rigorously established that it is selection consistent, oracle for parameter estimation, stable, and has a grouping property for highly correlated covariates. We illustrate through empirical studies that the BAR estimator has satisfactory performance for variable selection and parameter estimation. We have also extended the application of BAR to the sHDMSS domain by taking advantage of the fact that the BAR algorithm allows us to easily adapt existing high performance algorithms and software for massive ℓ_2-penalized Cox regression.[12]

Our surrogate ℓ_0-based BAR method and theory can be easily extended to a surrogate ℓ_d-based BAR method for any $d \in [0, 1]$, by replacing $\left(\hat{\beta}_j^{(k-1)}\right)^2$ with $\left|\hat{\beta}_j^{(k-1)}\right|^{2-d}$ in (4). We have observed empirically that, as d increases toward 1, the resulting estimator becomes less sparse, and the average number of FP as well as estimation bias tend to increase, especially for larger p_n, while the average number of FN tends to decrease. In practice, d can be used as a resolution tuning parameter.

Our theoretical and empirical results have established the BAR method as a valid and viable tool for variable selection and parameter estimation under the $p_n < n$ setting although p_n is allowed to diverge with n. Theoretical properties of the BAR estimator for the high-dimensional setting ($p_n \gg n$) remain to be investigated. Furthermore, as pointed out by a referee, although BAR is selection consistent and oracle, it is subject to the same postselection inference issues as other variable selection methods.[50,51] Lastly, although iteratively performing reweighted ℓ_2-penalizations allows BAR to enjoy the best of ℓ_0- and ℓ_2-penalized regressions and to readily adopt an existing efficient implementation of ℓ_2-penalization for sHDMSS data, its iterative nature does present another layer of computational complexity. While this added layer of computational complexity is not a practical concern for moderate size data, it can considerably increase the runtime in a large data setting when both n and p are large. As illustrated in our real data example, trying a prefixed tuning parameter value based on the extended BIC $\lambda_n = c\log(p_n)$ can reduce the runtime of BAR with reasonably good performance. To further improve its computational efficiency, we are currently developing some modified BAR algorithms including a cyclic coordinatewise BAR algorithm, which will have comparable computational complexity and runtime to other popular variable methods such as LASSO. This line of further developments is beyond the scope of this paper and will be fully studied in a sequel paper.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors are grateful to the referees for their insightful comments and suggestions that have greatly improved the paper. The authors are also grateful to Drs Randall Burd and Sushil Mittal for providing us access to the NTDB data. Gang Li's research was supported in part by National Institutes of Health grants P30 CA-16042, UL1TR000124-02, and P01AT003960.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the American College of Surgeons.[1] Restrictions apply to the availability of these data.

REFERENCES

1. National Trauma Data Bank. https://www.facs.org/quality-programs/trauma/tqp/center-programs/ntdb
2. Schuemie MJ, Ryan PB, Hripcsak D, Madigan G, Suchard MA. Honest learning for the healthcare system: large-scale evidence from real-world data. Science. 2017. Under review.
3. Tibshirani RThe lasso method for variable selection in the Cox model. Statist Med. 1997;16(4):385–395.

4. Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. Ann Stat. 2002;30(1):74–99.
5. Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. Biometrika. 2007;94(3):691–703.
6. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.
7. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1.
8. Johnson BA, Long Q, Huang Y, Chansky K, Redman M. Log-Penalized Least Squares, Iteratively Reweighted Lasso, and Variable Selection for Censored Lifetime Medical Cost. Technical ReportAtlanta, GA: Emory University; 2012.
9. Su X, Wijayasinghe CS, Fan J, Zhang Y. Sparse estimation of Cox proportional hazards models via approximated information criteria. Biometrics. 2016;72(3):751–759. [PubMed: 26873398]
10. Dai L, Chen K, Sun Z, Liu Z, Li G. Broken adaptive ridge regression and its asymptotic properties. J Multivar Anal. 2018;168:334–351. [PubMed: 30911202]
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–320.
12. Mittal S, Madigan D, Burd RS, Suchard MA. High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. Biostatistics. 2014;15(2):207–221. [PubMed: 24096388]
13. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.
14. Zou HThe adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429.
15. Akaike HA new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–723.
16. Schwarz GEstimating the dimension of a model. Ann Stat. 1978;6(2):461–464.
17. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. Biometrics. 2000;56(1):256–262. [PubMed: 10783804]
18. Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. J Am Stat Assoc. 2012;107(497):223–232. [PubMed: 22736876]
19. Breiman LHeuristics of instability and stabilization in model selection. Ann Stat. 1996;24(6):2350–2383.
20. Zhao H, Sun D, Li G, Sun J. Variable selection for recurrent event data with broken adaptive ridge regression. Can J Stat. 2018;46(3):416–428. 10.1002/cjs.11459 [PubMed: 32999527]
21. Zhao H, Wu Q, Li G, Sun J. Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. J Am Stat Assoc. 2019:1–13. 10.1080/01621459.2018.1537922 [PubMed: 34012183]
22. Zhao H, Sun D, Li G, Sun J. Simultaneous estimation and variable selection for incomplete event history studies. J Multivar Anal. 2019;171:350–361.
23. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–202.
24. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat. 1982;10(4):1100–1120.
25. Verweij PJ, Van Houwelingen HC. Penalized likelihood in Cox regression. Statist Med. 1994;13(23–24):2427–2436.
26. Frommlet F, Nuel G. An adaptive ridge procedure for L0 regularization. PLOS ONE. 2016;11(2):e0148620. [PubMed: 26849123]
27. Fan J, Feng Y, Wu Y. High-dimensional variable selection for Cox's proportional hazards model. In: Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. BrownBethesda, MD: Institute of Mathematical Statistics; 2010:70–86.
28. Yang G, Yu Y, Li R, Buu A. Feature screening in ultrahigh dimensional Cox's model. Statistica Sinica. 2016;26:881–901. [PubMed: 27418749]
29. Kawaguchi ES. Scalable Methods for Big Time-to-Event Data [PhD thesis]. Los Angeles, CA: UCLA; 2019.

30. Craven P, Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik (Heidelb). 1978;31(4):377–403.
31. Verweij PJ, Van Houwelingen HC. Cross-validation in survival analysis. Statist Med. 1993;12(24):2305–2314.
32. Ni A, Cai J. Tuning parameter selection in Cox proportional hazards model with a diverging number of parameters. Scand J Stat Theory Appl. 2018;45(3):557–570.
33. Wang H, Li R, Tsai C-L. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007;94(3):553–568. [PubMed: 19343105]
34. Zhang Y, Li R, Tsai C-L. Regularization parameter selections via generalized information criterion. J Am Stat Assoc. 2010;105(489):312–323. [PubMed: 20676354]
35. Zhang T, Oles FJ. Text categorization based on regularized linear classification methods. Information Retrieval. 2001;4(1):5–31.
36. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat. 2008;2(1):224–244.
37. Gorst-Rasmussen A, Scheike T. Coordinate descent methods for the penalized semiparametric additive hazards model. J Stat Softw. 2012;47(9):1–17. https://www.jstatsoft.org/v047/i09
38. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Trans Model Comput Simul. 2013;23(1).
39. Mittal S, Madigan D, Cheng JQ, Burd RS. Large-scale parametric survival analysis. Statist Med. 2013;32(23):3955–3971.
40. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. [PubMed: 20808728]
41. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat. 2011;5(1):232–253. [PubMed: 22081779]
42. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. 2007;49(3):291–304.
43. Zhang X, Cheng G. Simultaneous inference for high-dimensional linear models. J Am Stat Assoc. 2017;112(518):757–768.
44. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008;95(3):759–771.
45. Chen J, Chen Z. Extended BIC for small-n-large-P sparse GLM. Statistica Sinica. 2012;22(2):555–574.
46. Gao X, Carroll RJ. Data integration with high dimensionality. Biometrika. 2017;104(2):251–272. [PubMed: 28757650]
47. Yang Y, Zou H. A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. Stat Interface. 2012;6(2):167–173.
48. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–2546. [PubMed: 7069920]
49. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statist Med. 1996;15(4):361–387.
50. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. Ann Stat. 2014;42(2):413. [PubMed: 25574062]
51. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. Ann Stat. 2016;44(3):907–927.

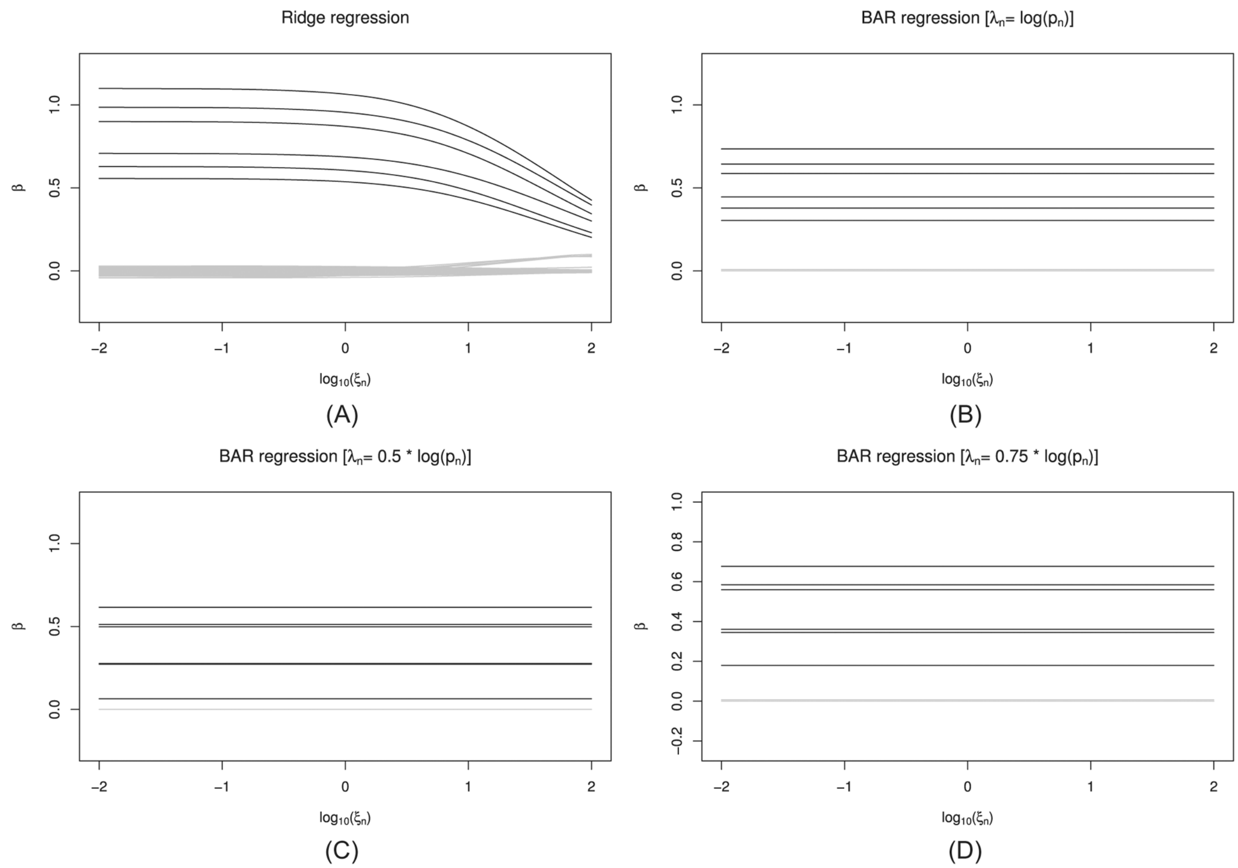

FIGURE 1.

Path plot for broken adaptive ridge (BAR) regression with varying (A) ξ_n and (B) $\lambda_n = \log(p_n)$, (C) $\lambda_n = 0.5 \log(p_n)$, and (D) $\lambda_n = 0.75 \log(p_n)$ with estimates averaged over 100 Monte Carlo simulations of size $n = 300$, $p_n = 100$, and censoring rate ≈ 25%. Path plot for ridge regression (D) with varying ξ_n is also included as a comparison

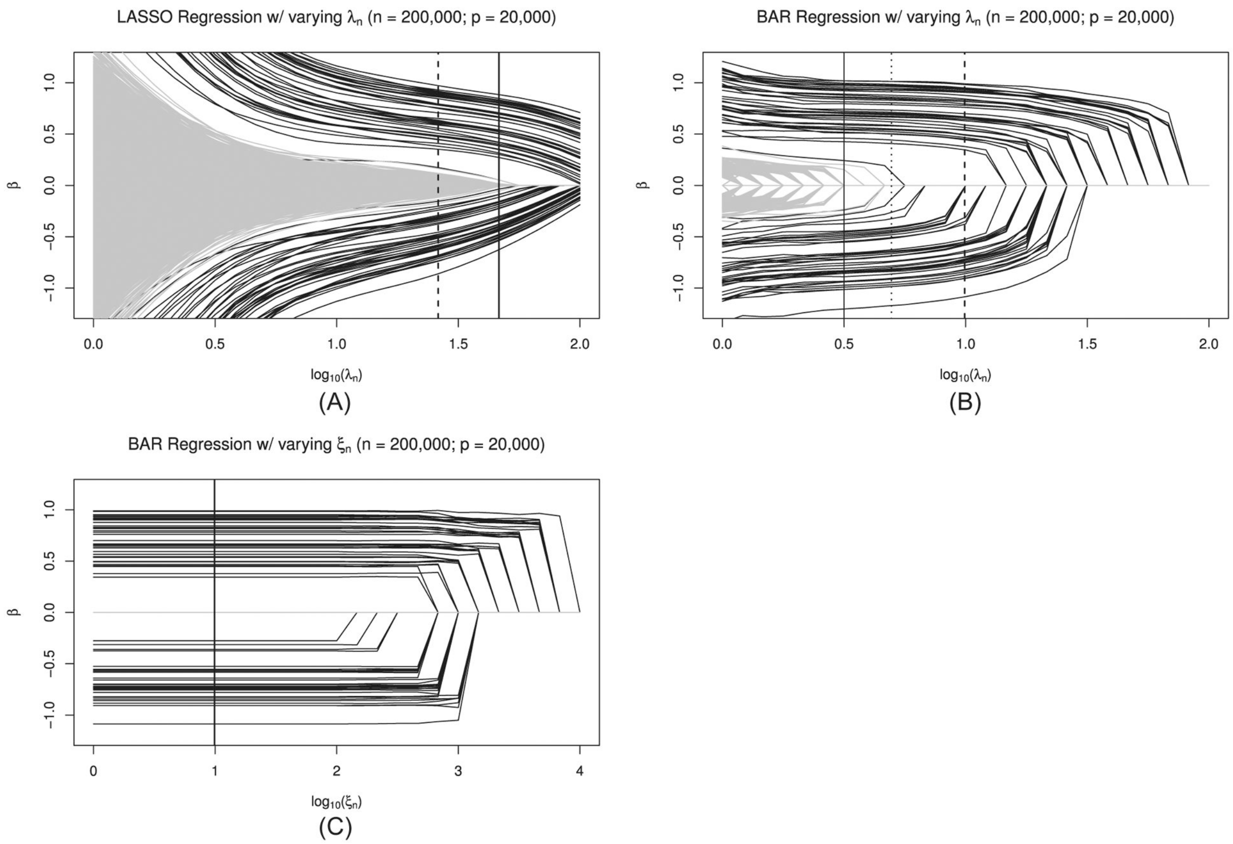

FIGURE 2.

Path plots for massive sparse Cox's regression for LASSO (mCox-LASSO) and massive Cox's regression for broken adaptive ridge (mBAR) regression. A, Path plot for mCox-LASSO regression, where the black solid and dashed lines represents the estimates when BIC minimization and cross-validation where used to find the optimal value of the tuning parameter, respectively; B, Path plot for mBAR regression with $\xi_n = \log(p_n)$ and varying λ_n, where the black solid, dashed, and dotted lines represent estimates where λ_n was found using Bayesian information criterion minimization, fixed at $\log(p_n)$ and $0.5 \log(p_n)$, respectively; C, Path plot for mBAR regression with $\lambda_n = \log(p_n)$ and varying ξ_n, where the black solid line represent the estimates for mBAR when $\xi_n = \log(p_n)$

TABLE 1

(Moderate dimension and sample size) Simulated estimation and variable selection performance of broken adaptive ridge (BAR) Bayesian information criterion (BIC), LASSO (BIC), SCAD (BIC), adaptive lasso (ALASSO) (BIC), and MCP (BIC) where BIC in parenthesis indicates that the BIC minimization was used to select the tuning parameters via a grid search. (MSB = mean squared bias; FN = mean number of false positives; FP = mean number of false negatives; SM = average similarity measure; BIC = average BIC score; Each entry is based on 100 Monte Carlo samples of size $n = 300, 1000$, $p_n = 100$, censoring rate = 25%)

		MSB	FN	FP	SM	BIC
$n = 300$	BAR ($\lambda_n = 0.5\log(p_n)$)	0.06	0.02	0.23	0.98	1930.97
	BAR ($\lambda_n = \log(p_n)$)	0.10	0.17	0.02	0.98	1938.43
	BAR (BIC)	0.11	0.01	1.79	0.89	1919.26
	LASSO (BIC)	0.27	0.01	3.32	0.82	1958.40
	SCAD (BIC)	0.12	0.01	2.23	0.87	1933.43
	ALASSO (BIC)	0.11	0.04	1.48	0.90	1935.60
	MCP (BIC)	0.09	0.02	1.21	0.92	1929.33
$n = 1000$	BAR ($\lambda_n = 0.5\log(p_n)$)	0.01	0.00	0.19	0.99	8200.97
	BAR ($\lambda_n = \log(p_n)$)	0.01	0.00	0.00	1.00	8203.52
	BAR (BIC)	0.02	0.00	0.73	0.95	8196.51
	LASSO (BIC)	0.10	0.00	2.77	0.84	8236.76
	SCAD (BIC)	0.01	0.00	0.23	0.98	8203.00
	ALASSO (BIC)	0.02	0.00	0.26	0.98	8204.58
	MCP (BIC)	0.01	0.00	0.08	0.99	8202.04

TABLE 2

(Sparse high-dimensional and massive sample size) Estimation and variable selection results for massive Cox regression with broken adaptive ridge (mBAR) and LASSO penalty (mCox-LASSO[12]) for a simulated sHDMSS data set with $n = 200\,000$, $p_n = 20\,000$, and $q_n = 80$. (Bias $= \|\hat{\beta} - \beta_0\|_2$; FP= number of false positives; FN = number of false negatives)

Method	Bias	FP	FN	BIC score
mBAR ($\lambda_n = 0.5\log(p_n)$)	1.19	0	3	83 313.02
mBAR ($\lambda_n = \log(p_n)$)	2.02	0	10	83 573.96
mBAR (BIC)	**0.97**	**5**	0	**83 266.47**
mCox-LASSO (BIC)	2.93	12	0	84 479.47
mCox-LASSO (CV)	2.12	963	0	93 770.58

Abbreviations: BIC, Bayesian information criterion; CV, cross-validation.

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 3

(Pediatric National Trauma Data Bank (NTDB) data) Comparison of mCox-LASSO and massive Cox's regression for broken adaptive ridge (mBAR) regression for the pediatric NTDB data. (mCox-LASSO cross-validation (CV) and mCox-LASSO Bayesian information criterion (BIC) correspond to mCox-LASSO using cross validation and BIC selection criterion, respectively. mBAR (BIC) denotes mBAR using the BIC selection criterion while fixing $\xi_n = \log(p_n)$. The training set has a sample size of 168 000, while the test set used for the *c*-index has a sample size of 45 555)

Method	# Selected	BIC score	*c*-index	Runtime (hours)
mBAR ($\lambda_n = 0.5 \log(p_n)$)	45	51 613.52	0.91	8
mBAR ($\lambda_n = \log(p_n)$)	21	52 182.90	0.89	8
mBAR (BIC)	83	51 269.43	0.93	97
mCox-LASSO (BIC)	100	52 544.90	0.91	25
mCox-LASSO (CV)	253	53 165.44	0.92	41