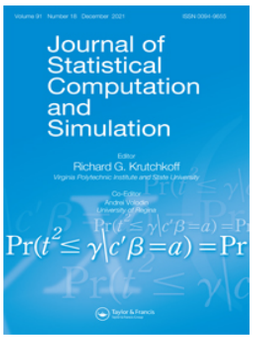

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gscs20

Global optimal model selection for high-dimensional survival analysis

Guotao Chu & Gyuhyeong Goh

To cite this article: Guotao Chu & Gyuhyeong Goh (2021) Global optimal model selection for high-dimensional survival analysis, Journal of Statistical Computation and Simulation, 91:18, 3850-3863, DOI: 10.1080/00949655.2021.1954183

To link to this article: https://doi.org/10.1080/00949655.2021.1954183

Published online: 18 Jul 2021.

Submit your article to this journal

Article views: 114

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gscs20

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2021, VOL. 91, NO. 18, 3850–3863
https://doi.org/10.1080/00949655.2021.1954183

Taylor & Francis
Taylor & Francis Group

Check for updates

Global optimal model selection for high-dimensional survival analysis

Guotao Chu and Gyuhyeong Goh

Department of Statistics, Kansas State University, Manhattan, KS, USA

ABSTRACT

With the popularity of high-dimensional data, model selection is of great importance in recent survival analysis. In a model selection context, an important research question is how to define the best model. To answer this, various model selection criteria have been proposed for defining the best model. The existing methods commonly use the L_0-norm penalization in order to measure the model complexity based on the number of parameters. However, due to the nonconvexity of the L_0-penalty, finding the best model via global optimization has been a challenging research subject in statistics and machine learning. In this paper, we propose a global optimization algorithm using a modification of the simulated annealing, which is a probabilistic search algorithm for the global optimum in statistical mechanics. The performance of the proposed method is examined via simulation study and real data analysis.

ARTICLE HISTORY
Received 9 November 2020
Accepted 7 July 2021

KEYWORDS
Boltzmann distribution; cox proportional hazard model; generalized information criterion; high-dimensional variable selection

1. Introduction

In high-dimensional survival analysis, a primary goal is to identify relevant covariates that are related to the survival time. Various model selection criteria have been developed using a L_0-norm penalty function. For example, the Bayesian information criterion (BIC) [1] and Akaike information criterion (AIC) [2] are the most popular choice for classical regression modelling. In a high-dimensional regression setting, Chen and Chen [3] propose a modified version of BIC, called the extended BIC, to consistently select the true data-generating model over large model spaces.

Since the use of L_0-penalty yields a nonconvex optimization problem, finding the best model, which is the global optimum of the model selection criterion, is computationally expensive and time-consuming in a high-dimensional data setting in which the number of candidate covariates is large. To reduce the heavy computational burden in a high-dimensional variable selection problem, penalized partial-likelihood estimation with a convex surrogate penalty has been proposed in the literature. For example, penalized likelihood estimation methods with convex penalties (e.g. lasso [4], adaptive lasso [5], and elastic net [6]) are developed for high-dimensional Cox proportional hazards regression [7–9]. However, these approaches do not ensure the convergence to the global optimum of

CONTACT Gyuhyeong Goh ggoh@ksu.edu

© 2021 Informa UK Limited, trading as Taylor & Francis Group

the model selection criterion since the solution path is generated by only a finite sequence of tuning parameters. In other words, due to the limited coverage of the tuning parameter values, there is a high chance that the solution path has missed the global solution to the model selection criterion.

In thermodynamics, Kirkpatrick et al. [10] propose a global optimization algorithm, called simulated annealing. The key idea of simulated annealing is to perform a stochastic search so that we can avoid the chance to get stuck in a local optimum. Using the Metropolis-Hastings sampling [11], the move of simulated annealing generates a Markov chain with a stationary distribution whose mode is the same as the global optimum of the target function. Although simulated annealing assures the convergence to the global optimum in a nonconvex optimization framework, the slow convergence and the choice of the proposal distribution are regarded as major drawbacks. As a result, the application of simulated annealing for high-dimensional variable selection is computationally infeasible since its computational efficiency drops dramatically as the number of covariates increases.

In this paper, we propose a new global optimization method for high-dimensional survival model selection with a general class of model selection criteria, often referred to as generalized information criterion [12–14]. The key idea of the proposed method is to incorporate Gibbs sampling into a simulated annealing framework via the concept of Boltzmann distribution in statistical mechanics. The proposed method enables us to perform a probabilistic search using the Gibbs sampler, which leads to the fast and stable convergence to the target distribution [15]. In addition, the use of the Gibbs sampler yields that the probability of accepting the move to a new model always becomes one so that it automatically eliminates the issue of proposal distribution selection in the traditional simulated annealing algorithm. The technical details of the proposed method are given in Section 4. As shown in Section 5, our proposed method outperforms many existing methods. The real data analysis in Section 6 also demonstrates the applicability of the proposed method to a blood cancer study.

2. Basic setup and generalized information criterion

For subject $i \in \{1, \ldots, n\}$, let $T_i = \min(T_i^*, C_i)$ be the observed failure time and $\mathbf{x}_i$ be the p-dimensional vector of possible covariates, where T_i^* is the actual death time of the i-th individual and C_i is the censoring time. Denote by $\delta_i = I\{T_i \leq C_i\}$ the indicator of the occurrence of the event, where $I\{\cdot\}$ represents an indicator function.

In survival analysis, the Cox proportional hazard model [16] is a widely used semiparametric regression model. The Cox model provides a way to examine how the covariates are associated with the rate of a particular event happening (e.g. death) at time t, where the rate is referred as the hazard rate. Specifically, the Cox model explains the relationship between the hazard function and the covariates by assuming the form

$$h(t) = h_0(t) \times \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}), \tag{1}$$

where $h(t)$ is the hazard function at time t, $h_0(t)$ is the baseline hazard function, which reflects the underlying hazard for subjects with all covariates equal to 0, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the p-dimensional coefficient vector, which is of our primary interest.

Under the right-censored scenario with observations $\{(T_i, \mathbf{x}_i, \delta_i), i = 1, \ldots, n\}$, the regression parameter $\boldsymbol{\beta}$ in the Cox model (1) can be estimated by constructing the partial likelihood without imposing a distributional assumption on the data. The partial log-likelihood function is defined by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log \left\{ \sum_{l \in R(T_i)} \exp(\mathbf{x}_l^{\mathrm{T}} \boldsymbol{\beta}) \right\} \right]. \tag{2}$$

where $R(T_i)$ is the risk set at time T_i, which represents the number of individuals who survived at least until time T_i.

Under a low-dimensional regression setting (i.e. $p \ll n$), it is well known that the asymptotic normality holds for the maximum partial likelihood estimator, which is obtained by maximizing (2). However, when the number of covariates p is large, variable selection is necessary to eliminate irrelevant covariates from the model so that the useful asymptotic property can be achieved under the reduced model.

In a high-dimensional Cox regression setting, the best model can be identified by minimizing the penalized partial log-likelihood as follows:

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta}),$$

where $\mathrm{pen}(\boldsymbol{\beta})$ is a penalty function which increases as the number of parameters increases. In the model selection literature, the penalty function is commonly assumed to be a linear function of L_0-norm, that is,

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0, \tag{3}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} I\{\beta_j \neq 0\}$ denotes the L_0-norm and λ is a prespecified tuning parameter controlling the degrees of penalization. The model selection criterion in (3) is often referred to as generalized information criterion (GIC). According to the choice of λ, GIC reduces to a well-known model selection criterion. For example, when we set $\lambda = \ln(n_0)$, GIC becomes BIC for censored survival models [17], where $n_0 = \sum_{i=1}^{n} \delta_i$.

Let $\boldsymbol{s} = (s_1, \ldots, s_p)$ represent a reduced Cox model such that $\beta_j \neq 0$ if $s_j = 1$ and $\beta_j = 0$ if $s_j = 0$ for $j = 1, \ldots, p$. Given $\boldsymbol{s}$, let $\boldsymbol{\beta}(\boldsymbol{s})$ be the sub-vector of $\boldsymbol{\beta}$ corresponding to one elements in $\boldsymbol{s}$. Then, the form of GIC in (3) can be further generalized as follows:

$$\mathrm{GIC}(\boldsymbol{s}) = -2l(\hat{\boldsymbol{\beta}}(\boldsymbol{s})) + \mathrm{pen}(|\boldsymbol{s}|), \tag{4}$$

where $|\boldsymbol{s}| = \sum_{j=1}^{p} s_j$ denotes the number of parameters under model $\boldsymbol{s}$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ is the maximum partial likelihood estimate of $\boldsymbol{\beta}(\boldsymbol{s})$ under model $\boldsymbol{s}$. Note that, when $\mathrm{pen}(|\boldsymbol{s}|) = \lambda|\boldsymbol{s}|$, (4) reduces to the original form of GIC in (3). Throughout this paper, GIC refers to the form of (4).

Due to the nonconvexity of the penalty function in (4), model selection with GIC should be performed by a brute-force algorithm (i.e. by comparing all possible models). However, when p is large, finding the best model using a brute-force algorithm, often called best subset selection in the statistical literature [18], becomes an NP-hard problem. For example, when $p = 100$, we need to compare $2^{100} \approx 1.27^{30}$ candidate models.

3. Convex surrogate

In the recent high-dimensional regression research, sparse estimation with convex penalties has been extensively studied. For example, using the L_1-norm penalty, Tibshirani [4] proposes the lasso (least absolute shrinkage and selection operator),

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1,$$

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta|_j$ and $\lambda \geq 0$. To improve the statistical efficiency of lasso, Zou [5] develops the adaptive lasso by using a weighted L_1-norm as follows:

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w_j|\beta_j|,$$

where $w_1, \ldots, w_p$ are data-driven weights. For high-dimensional and correlated data, Zou and Hastie [6] propose the elastic net,

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||^2,$$

where $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2$ and λ_1 and λ_2 are non-negative tuning parameters. Note that the elastic net includes the lasso as a special case with $\lambda_2 = 0$. Since these penalties leads to not only the convexity of the objective function but also the sparse estimates of $\boldsymbol{\beta}$, they have been considered as a solution to high-dimensional variable selection. However, unlike GIC, the tuning parameters are unknown in the penalized likelihood estimation framework and they must be chosen by a model selection criterion. From this aspect, the penalized likelihood estimation with GIC tuning parameter selection can be considered as a convex surrogate of the GIC model selection. This procedure can be summarized as follows:

- Step 1: Define a sequence of values for tuning parameter λ, Λ. For example, $\Lambda = \{\lambda_t = \epsilon(t-1) : t = 1, \ldots, T\}$ for small $\epsilon > 0$.
- Step 2: For each value of $\lambda \in \Lambda$, compute the sparse estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}(\lambda)$, by minimizing the penalized likelihood function given λ. For example, $\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$ for lasso.
- Step 3: Let $\boldsymbol{s}_\lambda = \{(s_{\lambda 1}, \ldots, s_{\lambda p}) : s_{\lambda j} = I\{\hat{\beta}_j(\lambda) \neq 0\}, j = 1, \ldots, p\}$ be the reduced model given $\hat{\boldsymbol{\beta}}(\lambda)$. Compute GIC($\boldsymbol{s}_\lambda$) for all $\lambda \in \Lambda$.
- Step 4: Find the best model by $\min_{\lambda \in \Lambda}$ GIC($\boldsymbol{s}_\lambda$).

Note that although the above procedure is computationally efficient and fast, there is a main limitation that the best model is generally a local optimum, not the global optimum because the solution path has been generated by a finite sequence of λ-values. In the following section, we introduce our proposed solution to global optimum model selection with GIC.

4. Global optimal model selection

In this section, we introduce a new method to find the global optimal model using GIC. Our proposed method is motivated by the idea of simulated annealing, which is a popular global optimization algorithm in statistical mechanics.

4.1. Simulated annealing

Simulated annealing (SA), originally proposed by Kirkpatrick et al. [10], is a stochastic optimization method for finding the global optimum in a non-convex optimization problem. The technique mimics the process of annealing in metallurgy, which is a technique involving heating and cooling of a material to increase the size of its crystals and reduce their defects. Let $E(\boldsymbol{s})$ be a energy function at state $\boldsymbol{s}$. In general, SA is used to find the state that leads to a global minimum energy.

In statistical thermodynamics, the probability of a physical system being in state $\boldsymbol{s}$ with energy $E(\boldsymbol{s})$ at temperature τ can be described by the Boltzmann distribution [19],

$$p_\tau(\boldsymbol{s}) \propto \exp\left\{-\frac{E(\boldsymbol{s})}{\kappa\tau}\right\}, \tag{5}$$

where κ is the Boltzmann's constant, which is usually a known constant in SA. Using the Boltzmann distribution, SA always converges to the global optimum by performing a stochastic search with annealing. The detailed SA algorithm is given as follows:

- Step 1: Set an initial state $\boldsymbol{s} = \boldsymbol{s}_0$ and an initial temperature $\tau = \tau_0$.
- Step 2:
  (a) Draw a new state $\boldsymbol{s}^*$ from a proposal distribution $q(\boldsymbol{s}^* \mid \boldsymbol{s})$, which represents the conditional probability of $\boldsymbol{s}^*$ given the current value of $\boldsymbol{s}$.
  (b) Move to the new state $\boldsymbol{s}^*$ with probability

$$\min\left\{1, \frac{p_\tau(\boldsymbol{s}^*)q(\boldsymbol{s} \mid \boldsymbol{s}^*)}{p_\tau(\boldsymbol{s})q(\boldsymbol{s}^* \mid \boldsymbol{s})}\right\}. \tag{6}$$

  (c) Repeat step (a) and step (b) until the chain is reached an equilibrium state.
- Step 3: Decrease the temperature by $\tau = \tau - \epsilon$ for small $\epsilon > 0$. If $\tau \leq 0$, then terminate. Otherwise, go to Step 2.

In the context of GIC optimization, we can employ SA by replacing the energy function $E(\boldsymbol{s})$ with $\mathrm{GIC}(\boldsymbol{s})$. Then, the Boltzmann distribution can be obtained by

$$p_\tau(\boldsymbol{s}) \propto \exp\left\{-\frac{\mathrm{GIC}(\boldsymbol{s})}{\kappa\tau}\right\}. \tag{7}$$

For the reminder of the paper, without loss of generality, we assume $\kappa = 2$. It is important to note that Step 2 in SA come from the idea of Metropolis-Hastings sampling, which is a Markov chain Monte Carlo (MCMC) method for sampling from a probability distribution when direct sampling is difficult. However, the specification of proposal distribution $q(\cdot \mid \cdot)$ is cumbersome in our setting. An even more serious problem is that the acceptance

probability in (6) tends to decrease exponentially as the number of covariates, p, increases. As a result, the probability of moving to a new state can be extremely small in a high-dimensional variable selection case.

To demonstrate, we perform a simulation study as follows: First, we generate artificial survival times of 100 subjects, $T_1^*, \ldots, T_{100}^*$, by $T_i^* = -\frac{\log(U_i)}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$, where $U_i \overset{iid}{\sim} \mathrm{Unif}(0,1)$, $\mathbf{x}_i \overset{iid}{\sim} \mathrm{N}_p(0,\Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$, and $\boldsymbol{\beta} = (0, -0.7, -0.7, 0, \ldots, 0)^{\mathrm{T}}$. Second, we create censoring time C_i by generating a sample from Exp(0.58), which produces about 40% censoring rate. Then, given the simulated data, we proceed the SA algorithm with $\tau = 1$ and count the acceptance rate of moving to a new state as follows:

- Step 1: Set $\boldsymbol{s} = (0, \ldots, 0)$, $\tau = 1$, and $a = 0$ (count for a new move).
- Step 2:
  (a) Generate $\boldsymbol{s}^* = (s_1^*, \ldots, s_p^*)$ with $s_j^* \overset{iid}{\sim} \mathrm{Ber}(0.5)$ for $j = 1, \ldots, p$.
  (b) Calculate $\Delta(\boldsymbol{s}, \boldsymbol{s}^*) = \exp\{-\frac{\mathrm{GIC}(\boldsymbol{s}^*) - \mathrm{GIC}(\boldsymbol{s})}{\kappa\tau}\}$, where GIC is chosen to be EBIC defined in (8).
  (c) Generate $u \sim \mathrm{Unif}(0,1)$.
  (d) If $\Delta(\boldsymbol{s}, \boldsymbol{s}^*) \geq u$, update $\boldsymbol{s} = \boldsymbol{s}^*$ and count $a = a + 1$. Otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.
- Step 3: Repeat Step 2 for 5000 times.

Figure 1 displays our simulation result. It clearly shows that the acceptance rate of moving to a new state, $\alpha = a/5000$, drops dramatically as p increases. In particular, when $p = 20$, there is almost no chance that the current state moves to a new state. In this case, SA cannot converge to the global optimum even if the iteration number is extremely large. To address the limitations of SA for high-dimensional variable selection, we propose a new stochastic search algorithm using the idea of Gibbs sampler. The details are discussed in the next section.

4.2. Proposed method

Motivated by the idea of Gibbs sampling, we propose to generate a candidate model for the next move in SA by adding a new predictor to or deleting one from the current model. To this end, let $\boldsymbol{s} = (s_1, \ldots, s_p)$ be the current state. For a given j, we define a candidate model by $\boldsymbol{s}^* = (s_1^*, \ldots, s_p^*)$ such that $s_k^* = s_k$ if $k \neq j$. Then, we can obtain the following important property.

Lemma 4.1: *Let $p_\tau(\cdot)$ be the Boltzmann distribution defined in* (7)*. Assume that the proposal distribution, $q(\boldsymbol{s}^* \mid \boldsymbol{s})$, in SA is proportional to $p_\tau(\boldsymbol{s}^*)$ with respect to s_j^*, that is, $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(\boldsymbol{s}_j^*)$ with respect to s_j^*. Then, in Step 2 of SA, $\boldsymbol{s}^*$ is accepted with probability one.*

Proof of Lemma 4.1: Since $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(\boldsymbol{s}_j^*)$ with respect to s_j^*, it can be viewed as $q(\boldsymbol{s}^* \mid \boldsymbol{s}) \propto p_\tau(s_j^* \mid \boldsymbol{s}_{-j}^*)$, where $\boldsymbol{s}_{-j}$ is obtained by deleting the j-th component of $\boldsymbol{s}$. Recall that by the definition of $\boldsymbol{s}^*$, we have $\boldsymbol{s}_{-j}^* = \boldsymbol{s}_{-j}$. Then, it follows that $q(\boldsymbol{s}^* \mid \boldsymbol{s}) = p_\tau(s_j^* \mid \boldsymbol{s}_{-j}) = p_\tau(s_j^* \mid \boldsymbol{s}_{-j}^*)$. Similarly, it can be shown that $q(\boldsymbol{s} \mid \boldsymbol{s}^*) = p_\tau(s_j \mid \boldsymbol{s}_{-j}^*) = p_\tau(s_j \mid \boldsymbol{s}_{-j})$. This

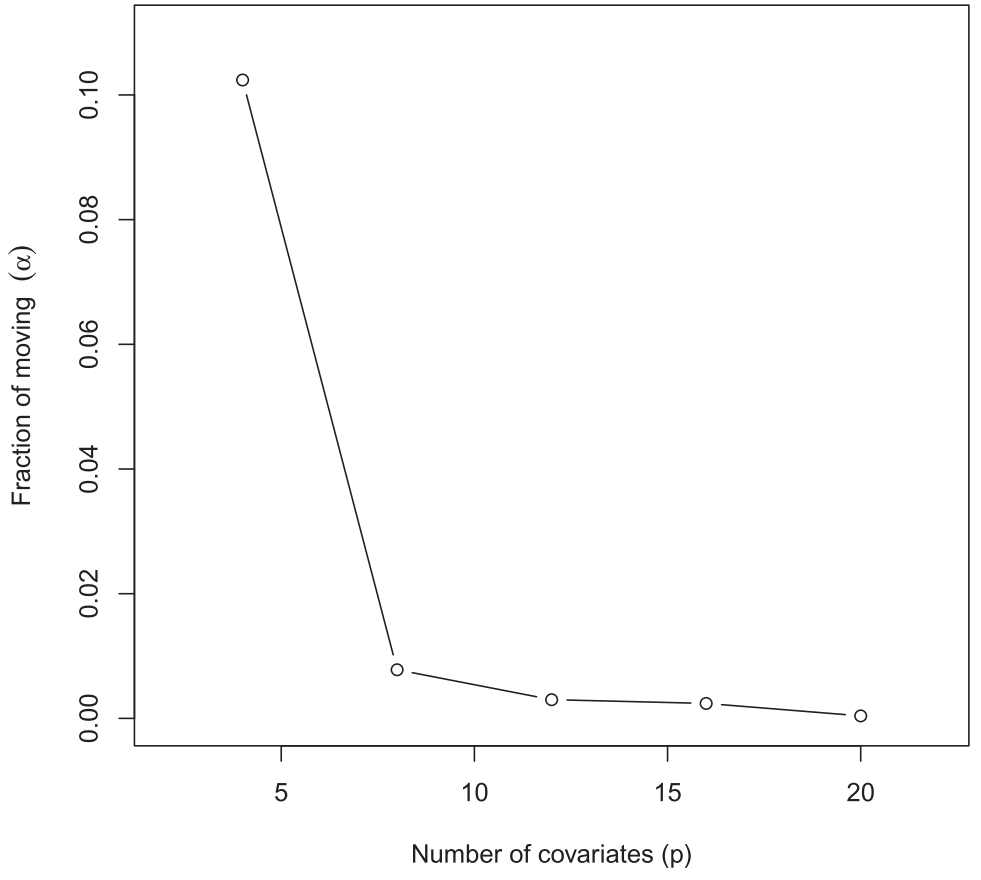

Figure 1. Changes in the acceptance rate of moving to a new state in SA when the number of covariates p increases.

implies that

$$\begin{aligned}\frac{p_\tau(\boldsymbol{s}^*)q(\boldsymbol{s} \mid \boldsymbol{s}^*)}{p_\tau(\boldsymbol{s})q(\boldsymbol{s}^* \mid \boldsymbol{s})} &= \frac{p_\tau(\boldsymbol{s}^*)p_\tau(s_j \mid \boldsymbol{s}^*_{-j})}{p_\tau(\boldsymbol{s})p_\tau(s^*_j \mid \boldsymbol{s}_{-j})} \\ &= \frac{p_\tau(\boldsymbol{s}^*)p_\tau(s_j \mid \boldsymbol{s}_{-j})}{p_\tau(\boldsymbol{s})p_\tau(s^*_j \mid \boldsymbol{s}^*_{-j})} \\ &= \frac{p_\tau(\boldsymbol{s}^*_{-j})}{p_\tau(\boldsymbol{s}_{-j})} \\ &= 1,\end{aligned}$$

where the last equality holds from the fact that $\boldsymbol{s}_{-j} = \boldsymbol{s}^*_{-j}$. This completes our proof. ■

Now, we define the proposal distribution by

$$q(\boldsymbol{s}^* \mid \boldsymbol{s}) = \frac{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(\boldsymbol{s}^*)\right\} I\left\{\boldsymbol{s}^*_{-j} = \boldsymbol{s}_{-j}\right\}}{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(s^*_j = 1, \boldsymbol{s}^*_{-j})\right\} + \exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(s^*_j = 0, \boldsymbol{s}^*_{-j})\right\}}.$$

In SA, one of key features is the annealing process, that is, the temperature τ decreases as the iteration number increases. Unlike SA, in our proposed method, we consider to increase the temperature for the following reason. Under the given temperature, the selected best model can be either a global optimum or a local optimum. If the best model is obtained at a local optimum, increasing the temperature will improve the chance to get out from the local trap and move forward to the global optimum. If the current best model is attained at the global optimum, then the current best model continuously remains the same as τ increases. Hence, in this case, we can conclude the convergence to the global optimum. The proposed algorithm is given in Algorithm 1.

Algorithm 1 Proposed algorithm.

Start from an initial state of $\boldsymbol{s} = (s_1, \ldots, s_p)$ with $\tau = \tau_0$, use $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_p)$ to store the best model, set $r = 1$, and define k to control the maximum model size. The algorithm proceeds as follows:

- Step 1: For $j = 1, \ldots, p$, update s_j by repeating the following steps until $r \leq pm$, where m is set to control the number of iterations.
  (a) Define $\boldsymbol{s}^*$ by $s_j^* = 1 - s_j$ and $s_i^* = s_j$ for $i \neq j$.
  (b) If $\sum_{i=1}^{p} s_i^* > k$, skip Steps (c)- -(d) below and jump to the next update for $j + 1$. Otherwise, calculate $\mathrm{GIC}(\boldsymbol{s}^*)$.
  (c) If $\mathrm{GIC}(\boldsymbol{s}^*) < \mathrm{GIC}(\hat{\boldsymbol{s}})$, update $\hat{\mathbf{s}} = \boldsymbol{s}^*$ and reset $r = 1$ and $\tau = \tau_0$. Otherwise, set $r = r + 1$.
  (d) We update $\boldsymbol{s} = \boldsymbol{s}^*$ if we obtain 1 from a Bernoulli trial with the success probability

$$\omega = \frac{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(\boldsymbol{s}^*)\right\}}{\exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(\boldsymbol{s}^*)\right\} + \exp\left\{-\frac{1}{\kappa\tau}\mathrm{GIC}(\boldsymbol{s})\right\}}.$$

  Otherwise, stay $\boldsymbol{s} = \boldsymbol{s}$.
- Step 2: Repeat Step 1 with $\boldsymbol{s} = \hat{\mathbf{s}}$, $r = 1$, and $\tau = \tau_{t+1} (> \tau_t)$ until $\tau = T_{\max}$, where $T_{\max}$ is a prespecified maximum temperature.

In Appendix, we provide a demo `R` code for implementing Step 1 of the proposed algorithm when GIC is assumed to be BIC.

5. Simulation study

In this section, we conduct a simulation study to investigate the performance of our proposed method under the Cox proportional hazard model. For the choice of GIC, we select the extended BIC (EBIC), which is the most popular choice for high-dimensional model selection [e.g. 20–22]:

$$\mathrm{EBIC}(\boldsymbol{s}) = -2l\left(\hat{\boldsymbol{\beta}}(\boldsymbol{s})\right) + \log(n_0)|\boldsymbol{s}| + 2\gamma\binom{p}{|\boldsymbol{s}|}, \tag{8}$$

where $\gamma \in [0, 1]$ is a pre-specified tuning parameter and we set $\gamma = 1$ so that EBIC always satisfies model selection consistency even when $p > n$ [3].

When we implement the proposed method using Algorithm 1, we set $\boldsymbol{s} = (0, 0, \ldots, 0)$, $k = 15$, $\tau \in \{1, 4/3\}$, and $m = 10$ for the initial setting, called 'proposed method (null) '. To perform sensitivity analysis in the setting of the initial estimate of $\boldsymbol{s}$, we also consider a random estimate of $\boldsymbol{s}$ by randomly choosing six elements of $\boldsymbol{s}$ to be one and setting the remaining elements to be zero, called 'proposed method (random) '. For the purpose of comparison, we employ the following four methods that are commonly used in high-dimensional variable selection: (1) lasso, (2) SCAD [23], (3) MCP [24], and (4) elastic net. The simulation study is conduced by using R, where lasso and elastic net are implemented by the `glmnet` package, and MCP and SCAD are implemented by the `ncvreg` package. For lasso, SCAD, MCP, and elastic net, we use the convex surrogate approach given in Section 3 with a grid of tuning parameters that are generated by the `R` packages.

We generate survival time T_i^* and censoring time C_i for $i = 1, \ldots, n$ independently as follows:

- $T_i^* = -\frac{\log(U_i)}{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$, where $U_i \overset{iid}{\sim} \mathrm{Unif}(0, 1)$, $\mathbf{x}_i \overset{iid}{\sim} \mathrm{N}_p(0, \Sigma)$ with $\Sigma = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$, and $\boldsymbol{\beta}$ is the p-dimensional vector with $\beta_1 = \beta_9 = 0.8$, $\beta_4 = \beta_{12} = -0.7$, $\beta_5 = \beta_{13} = 0.6$, and $\beta_j = 0$ for $j \neq 1, 4, 5, 9, 12, 13$.
- $C_i \overset{iid}{\sim} \mathrm{Exp}(\eta)$, where $\eta = 0.22$ (about 25% censoring rate) or $\eta = 0.57$ (about 40% censoring rate).

To consider various high-dimensional data settings, we consider the following 12 scenarios in the data-generating process:

(1) $n = 200, p = 100$, and censor rate= 25% ($\eta = 0.22$).
(2) $n = 200, p = 1000$, and censor rate= 25% ($\eta = 0.22$).
(3) $n = 200, p = 2000$, and censor rate= 25% ($\eta = 0.22$).
(4) $n = 500, p = 100$, and censor rate= 25% ($\eta = 0.22$).
(5) $n = 500, p = 1000$, and censor rate= 25% ($\eta = 0.22$).
(6) $n = 500, p = 2000$, and censor rate= 25% ($\eta = 0.22$).
(7) $n = 200, p = 100$, and censor rate= 40% ($\eta = 0.57$).
(8) $n = 200, p = 1000$, and censor rate= 40% ($\eta = 0.57$).
(9) $n = 200, p = 2000$, and censor rate= 40% ($\eta = 0.57$).
(10) $n = 500, p = 100$, and censor rate= 40% ($\eta = 0.57$).
(11) $n = 500, p = 1000$, and censor rate= 40% ($\eta = 0.57$).
(12) $n = 500, p = 2000$, and censor rate= 40% ($\eta = 0.57$).

To evaluate the performance of finding the global optimum model, for each method, we count the number of cases in which the EBIC evaluated at the optimal model is smaller than the other methods over 100 Monte Carlo replications. We denote by $\mathrm{F}_{\min}$ the ratio of finding the smallest EBIC out of the 100 replications. In addition, to access the variable performance, we calculate the false-positive rate (FPR) and the false negative

rate (FNR),

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad \text{and} \quad \text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}},$$

where TP, FP, TN and FN denote the number of true non-zeros, false non-zeros, true zeros, and false zeros, respectively.

The simulation result is summarized in Tables 1 and 2, where Time represents the average execution time in minutes over 30 replications when a Windows 10 computer with an Intel Core i7-8650U processor and 16 GB of memory is used. The result clearly shows that our proposed method always has the highest frequency of finding the smallest EBIC comparing with other methods for all 12 scenarios. In addition, the proposed method is less sensitive to the choice of the initial estimate of $\boldsymbol{s}$. This implies that our proposed method successfully identifies the global optimal model for the EBIC model selection procedure. It is also worth noting that the proposed method always achieves the lowest level of FPR and

Table 1. Simulation result with censoring rate = 25%.

(n, p)	Method	F_{min}	FPR	FNR	Time
(200,100)	Proposed method (null)	0.97	0.0022 (0.0006)	0.0050 (0.0037)	0.6997
	Proposed method (random)	0.98	0.0023 (0.0006)	0.0050 (0.0037)	0.7301
	LASSO	0.17	0.0182 (0.0016)	0.0867 (0.0163)	0.0123
	SCAD	0.25	0.0147 (0.0015)	0.0650 (0.0148)	0.0066
	MCP	0.61	0.0067 (0.0009)	0.0333 (0.0113)	0.0071
	Elastic Net	0.08	0.0197 (0.0018)	0.1517 (0.0212)	0.0130
(200,1000)	Proposed method (null)	0.84	0.0025 (0.0007)	0.1167 (0.0235)	7.4433
	Proposed method (random)	0.87	0.0030 (0.0007)	0.0900 (0.0193)	8.2580
	LASSO	0.05	0.0039 (0.0008)	0.5300 (0.0193)	0.0488
	SCAD	0.05	0.0048 (0.0010)	0.5183 (0.0204)	0.0662
	MCP	0.13	0.0111 (0.0016)	0.3167 (0.0293)	0.0434
	Elastic Net	0.05	0.0033 (0.0008)	0.5767 (0.0160)	0.0587
(200,2000)	Proposed method (null)	0.85	0.0011 (0.0003)	0.2133 (0.0288)	14.7546
	Proposed method (random)	0.86	0.0014 (0.0004)	0.1650 (0.0261)	15.1516
	LASSO	0.13	0.0016 (0.0005)	0.6217 (0.0148)	0.0226
	SCAD	0.14	0.0021 (0.0006)	0.6150 (0.0162)	0.0296
	MCP	0.15	0.0108 (0.0017)	0.4400 (0.0279)	0.0172
	Elastic Net	0.12	0.0013 (0.0004)	0.6350 (0.0120)	0.0266
(500,100)	Proposed method (null)	1.00	0.0013 (0.0005)	0.0000 (0.0000)	1.3108
	Proposed method (random)	1.00	0.0013 (0.0005)	0.0000 (0.0000)	1.3460
	LASSO	0.87	0.0015 (0.0005)	0.0000 (0.0000)	0.0448
	SCAD	0.96	0.0011 (0.0004)	0.0000 (0.0000)	0.0299
	MCP	0.99	0.0010 (0.0003)	0.0000 (0.0000)	0.0305
	Elastic Net	0.63	0.0049 (0.0008)	0.0000 (0.0000)	0.0456
(500,1000)	Proposed method (null)	1.00	0.0011 (0.0003)	0.0000 (0.0000)	12.1821
	Proposed method (random)	1.00	0.0011 (0.0003)	0.0000 (0.0000)	12.4426
	LASSO	0.68	0.0040 (0.0008)	0.0017 (0.0017)	0.1033
	SCAD	0.87	0.0018 (0.0005)	0.0000 (0.0000)	0.1783
	MCP	0.99	0.0012 (0.0004)	0.0000 (0.0000)	0.1291
	Elastic Net	0.51	0.0094 (0.0012)	0.0033 (0.0023)	0.1222
(500,2000)	Proposed method (null)	1.00	0.0014 (0.0004)	0.0000 (0.0000)	28.1613
	Proposed method (random)	1.00	0.0014 (0.0004)	0.0000 (0.0000)	29.5035
	LASSO	0.57	0.0064 (0.0011)	0.0100 (0.0052)	0.1151
	SCAD	0.70	0.0050 (0.0010)	0.0017 (0.0017)	0.1149
	MCP	0.96	0.0019 (0.0004)	0.0000 (0.0000)	0.0726
	Elastic Net	0.32	0.0129 (0.0015)	0.0017 (0.0065)	0.1320

Table 2. Simulation result with censoring rate = 40%.

(n, p)	Method	F_{min}	FPR	FNR	Time
(200,100)	Proposed method (null)	0.95	0.0018 (0.0005)	0.0383 (0.0120)	0.8466
	Proposed method (random)	0.98	0.0022 (0.0006)	0.0350 (0.0104)	0.8786
	LASSO	0.08	0.0139 (0.0015)	0.2183 (0.0232)	0.0210
	SCAD	0.16	0.0171 (0.0018)	0.1367 (0.0212)	0.0121
	MCP	0.40	0.0104 (0.0013)	0.0683 (0.0148)	0.0128
	Elastic Net	0.06	0.0143 (0.0017)	0.3000 (0.0248)	0.0224
(200,1000)	Proposed method (null)	0.79	0.0035 (0.0008)	0.2417 (0.0292)	7.2024
	Proposed method (random)	0.81	0.0040 (0.0010)	0.2333 (0.0284)	8.1090
	LASSO	0.18	0.0020 (0.0005)	0.6317 (0.0145)	0.0461
	SCAD	0.18	0.0027 (0.0007)	0.6200 (0.0158)	0.0538
	MCP	0.19	0.0078 (0.0013)	0.4633 (0.0275)	0.0354
	Elastic Net	0.18	0.0015 (0.0004)	0.6483 (0.0123)	0.0519
(200,2000)	Proposed method (null)	0.76	0.0025 (0.0006)	0.3800 (0.0304)	13.0769
	Proposed method (random)	0.89	0.0029 (0.0007)	0.2717 (0.0289)	14.4596
	LASSO	0.23	0.0017 (0.0004)	0.6633 (0.0103)	0.0232
	SCAD	0.24	0.0018 (0.0004)	0.6600 (0.0108)	0.0275
	MCP	0.24	0.0049 (0.0011)	0.5917 (0.0196)	0.0161
	Elastic Net	0.23	0.0011 (0.0003)	0.6683 (0.0102)	0.0275
(500,100)	Proposed method (null)	1.00	0.0013 (0.0004)	0.0000 (0.0000)	1.0255
	Proposed method (random)	1.00	0.0013 (0.0004)	0.0000 (0.0000)	1.1192
	LASSO	0.78	0.0033 (0.0006)	0.0000 (0.0000)	0.0448
	SCAD	0.96	0.0015 (0.0004)	0.0000 (0.0000)	0.0237
	MCP	1.00	0.0013 (0.0004)	0.0000 (0.0000)	0.0247
	Elastic Net	0.58	0.0069 (0.0011)	0.0000 (0.0000)	0.0457
(500,1000)	Proposed method (null)	1.00	0.0017 (0.0004)	0.0000 (0.0000)	12.6398
	Proposed method (random)	1.00	0.0017 (0.0004)	0.0000 (0.0000)	13.6700
	LASSO	0.46	0.0079 (0.0010)	0.0015 (0.0075)	0.0151
	SCAD	0.62	0.0061 (0.0010)	0.0017 (0.0017)	0.1572
	MCP	0.97	0.0020 (0.0005)	0.0000 (0.0000)	0.1128
	Elastic Net	0.36	0.0121 (0.0013)	0.0267 (0.0088)	0.1264
(500,2000)	Proposed method (null)	0.99	0.0013 (0.0006)	0.0000 (0.0000)	24.3697
	Proposed method (random)	0.98	0.0015 (0.0007)	0.0000 (0.0000)	26.6497
	LASSO	0.30	0.0135 (0.0014)	0.0250 (0.0080)	0.1195
	SCAD	0.43	0.0115 (0.0015)	0.0167 (0.0060)	0.1054
	MCP	0.82	0.0040 (0.0008)	0.0017 (0.0017)	0.0663
	Elastic Net	0.13	0.0163 (0.0017)	0.0583 (0.0120)	0.1388

FNR for all 12 scenarios. This means that the proposed method provides the best performance in identifying the true model for high-dimensional variable selection. As mentioned earlier, EBIC possesses model selection consistency, that is, the global optimum model tends to be the true model with high probability when the sample size is large. Hence, the superiority of our proposed method in model selection can be regarded as another evidence to demonstrate that the proposed method successfully finds the global optimal model in terms of EBIC.

6. Real data application

In this section, we conduct real data analysis with the Diffuse Large B-Cell Lymphoma (DLBCL) data [25,26]. DLBCL has been known to be the most common type of non-Hodgkin lymphoma in the United States and worldwide [27]. The dataset used this analysis is publicly available at the R package `ROC632`. The data contain information about 240

Table 3. Real data analysis result with DLBCL data.

	Index set of selected genes	EBIC
Proposed method	260,663,1127,1162	1307.769
Lasso	705	1322.035
SCAD	705	1322.035
MCP	260,705,867	1318.424
Elastic net	705	1322.035

DLBCL patients with the 7399 gene expressions. Since 5 observations have survival time equal to 0, we eliminate them and use the information of the remaining 235 patients for our analysis, where the censoring rate is 0.434.

First, we perform a pre-screening procedure to screen out redundant covariates that are obviously unrelated to the survival time. For every single convariate, we obtain p-value by fitting the Cox model with the single covariate and then exclude it from the analysis if the obtained p-value is greater than 0.05. After the screening procedure, 1163 genes are finally selected for our analysis. Then, we apply the proposed method and the existing methods (lasso, SCAD, MCP, and elastic net) as in Section 5. Table 3 displays our analysis result. The result shows that our proposed method provides the smaller EBIC ($=1307.769$). This implies that the model selected by our proposed method receives the strongest support from the observed data.

7. Conclusion remarks

We have proposed a global optimal model selection procedure with GIC using the notion of statistical mechanics. The superiority of the proposed method in high-dimensional variable selection has been shown by the simulation study and real data analysis.

While we have restricted our attention to the Cox model in this paper, the proposed method can be easily adapted to different parametric and semi-parametric survival models by replacing the partial likelihood with the likelihood or pseudo-likelihood functions. In addition, various choices of GIC can be considered in the proposed framework. For recent developments in model selection criterion that belongs to GIC, see Kim and Jeon [28] and references therein.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–464.
[2] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19(6):716–723.
[3] Chen J, Chen Z. Extended bayesian information criteria for model selection with large model spaces. Biometrika. 2008;95(3):759–771.
[4] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol). 1996;58(1):267–288.
[5] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429.
[6] 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.

[7] Tibshirani R. The Lasso method for variable selection in the cox model. Stat Med. 1997;16(4):385–395.
[8] Zhang HH, Lu W. Adaptive lasso for cox's proportional hazards model. Biometrika. 2007;94(3):691–703.
[9] Goeman JJ. L1 penalized estimation in the cox proportional hazards model. Biometrical J. 2010;52(1):70–84.
[10] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–680.
[11] Metropolis N, Rosenbluth AW, Rosenbluth MN, et al. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6):1087–1092.
[12] Atkinson AC. A note on the generalized information criterion for choice of a model. Biometrika. 1980;67(2):413–418.
[13] Zhang Y, Li R, Tsai CL. Regularization parameter selections via generalized information criterion. J Am Stat Assoc. 2010;105(489):312–323.
[14] Kim Y, Kwon S, Choi H. Consistent model selection criteria on high dimensions. J Machine Learning Res. 2012;13:1037–1057.
[15] Casella G, George EI. Explaining the gibbs sampler. Am Stat. 1992;46(3):167–174.
[16] Cox DR. Regression models and life-tables. J R Stat Soc: Ser B (Methodol). 1972;34(2):187–202.
[17] Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. Biometrics. 2000;56(1):256–262.
[18] James G, Witten D, Hastie T, et al. An introduction to statistical learning. Vol. 112, New York: Springer; 2013.
[19] Gibbs JW. Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics. New York: C. Scribner's sons; 1902.
[20] Chen J, Chen Z. Extended bic for small-n-large-p sparse glm. Stat Sin. 2012;22(2):555–574.
[21] Luo S, Xu J, Chen Z. Extended bayesian information criterion in the cox model with a high-dimensional feature space. Ann Inst Stat Math. 2015;67(2):287–311.
[22] Foygel R, Drton M. Extended bayesian information criteria for gaussian graphical models. Advances in neural information processing systems. 2010. p. 604–612.
[23] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.
[24] Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.
[25] Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N Eng J Med. 2002;346(25):1937–1947.
[26] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503–511.
[27] Lymphoma Research Foundation. Diffuse large B-cell lymphoma. 2020. Available from: https://lymphoma.org/aboutlymphoma/nhl/dlbcl/.
[28] Kim Y, Jeon JJ. Consistent model selection criteria for quadratically supported risks. Ann Stat. 2016;44(6):2467–2496.

Appendix. R code

```
# y: survival time # status: censoring indicator # x: covariates with
dimension n by p # tau: temperature # k: upper bound for the candidate model
size n0 <- sum(status) BIC<-function(s) {
 if(sum(s) == 0)){
  fit0  <- coxph(Surv(y, status!=0) ~ 1);
  return(-2*fit0$loglik)
 }else{
  fit <- coxph(Surv(time, status) ~ x[,s]);
  return(-2*fit$loglik[2]+log(n0)* length(fit$coef))
 }
```

```
} s <- rep(0,ncol(x)) best.s <- s best.BIC <- BIC(s) r <- 1 while(r < m*p)
 {
   s.prev <- s
   for (j in 1:ncol(x))
   {
    if( length(which(s==1)) < k ) {
      s.1  <- s
      s.1[j] <- 1
      s.0  <- s
      s.0[j] <- 0
      BIC1  <- BIC(which(s.1!=0))
      BIC0  <- BIC(which(s.0!=0))
      BIC.min<- min(BIC1,BIC0)
      s.min <- ifelse(BIC1 > BIC0, s.0, s.1)
      if( best.BIC > BIC.min ){
        r <- 1
        best.s  <- s.min
        best.BIC <- BIC.min
      }else{
        r <- r+1
      }
      w   <- exp(-BIC1/tau+BIC0/tau)
      prob  <- w/(1+w)
      s[j]  <- rbinom(1, 1, prob=prob)
    }else { next }
   }
}
```