Statistica Sinica Preprint No: SS-14-171R1	
Title	Feature screening in ultrahigh dimensional cox's model
Manuscript ID	SS-14-171
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2014.171
Complete List of Authors	Guangren Yang Ye Yu Runze Li and Anne Buu
Corresponding Author	Runze Li
E-mail	rzli@psu.edu

where u is a scaling constant to be specified and W is a diagonal matrix. Throughout this paper, we use $W = \text{diag}\{-\ell_p''(\boldsymbol{\beta})\}$, the matrix consisting of the diagonal elements of $-\ell_p''(\boldsymbol{\beta})$. This implies that we approximate $\ell_p''(\boldsymbol{\beta})$ by $u\,\text{diag}\{\ell_p''(\boldsymbol{\beta})\}$.

Remark. Xu and Chen (2014) proposed a feature screening procedure by iterative hard-thresholding algorithm (IHT) for generalized linear models with independently and identically distributed (iid) observations. They approximated the likelihood function $\ell(\boldsymbol{\gamma})$ of the observed data by a linear approximation $\ell(\boldsymbol{\beta})+(\boldsymbol{\gamma}-\boldsymbol{\beta})^T\ell'(\boldsymbol{\beta})$, but they also introduced a regularization term $-u\|\boldsymbol{\gamma}-\boldsymbol{\beta}\|^2$. Thus, the $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ in Xu and Chen (2014) would coincide with the one in (2.4) if one set $W = I_p$, the $p \times p$ identity matrix, but the motivation of our proposal indeed is different from theirs, and the working matrix W is not set to be I_p throughout this paper.

It can be seen that $g(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$, and under some conditions, $g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Since W is a diagonal matrix, $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \leq m \tag{2.5}$$

for given $\boldsymbol{\beta}$ and m. Note that the maximizer of $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1}W^{-1}\ell_p'(\boldsymbol{\beta})$. Denote $r_j = w_j\tilde{\gamma}_j^2$ with w_j being the j-th diagonal element of W for $j = 1, \cdots, p$, and sort r_j so that $|r_{(1)}| \geq |r_{(2)}| \geq \cdots \geq |r_{(p)}|$. The solution of maximization problem (2.5) is the hard-thresholding rule defined below

$$\widehat{\gamma}_j = \tilde{\gamma}_j I\{|r_j| > |r_{(m+1)}|\} \widehat{=} H(\tilde{\gamma}_j; m). \tag{2.6}$$

This enables us to effectively screen features by using the following algorithm:

Step 1. Set the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.

Step 2. Set $t = 0, 1, 2, \cdots$ and iteratively conduct Step 2a and Step 2b below until the algorithm converges.

Step 2a. Calculate $\tilde{\boldsymbol{\gamma}}^{(t)} = (\tilde{\gamma}_1^{(t)}, \cdots, \tilde{\gamma}_p^{(t)})^T = \boldsymbol{\beta}^{(t)} + u_t^{-1}W^{-1}(\boldsymbol{\beta}^{(t)})\ell_p'(\boldsymbol{\beta}^{(t)})$, and

$$\tilde{\boldsymbol{\beta}}^{(t)} = (H(\tilde{\gamma}_1^{(t)}; m), \cdots, H(\tilde{\gamma}_p^{(t)}; m))^T \widehat{=} \mathbf{H}(\tilde{\boldsymbol{\gamma}}^{(t)}; m). \tag{2.7}$$

Set $S_t = \{j : \tilde{\beta}_j^{(t)} \neq 0\}$, the nonzero index of $\tilde{\boldsymbol{\beta}}^{(t)}$.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\beta_1^{(t+1)}, \cdots, \beta_p^{(t+1)})^T$ as follows. If $j \notin S_t$, set $\beta_j^{(t+1)} = 0$; otherwise, set $\{\beta_j^{(t+1)} : j \in S_t\}$ be the maximum partial

likelihood estimate of the submodel S_t.

Unlike the screening procedures based on marginal partial likelihood methods proposed in Fan, Feng and Wu (2010) and further studied in Zhao and Li (2012), our proposed procedure is to iteratively update $\boldsymbol{\beta}$ using Step 2. This enables the proposed screening procedure to incorporate correlation information among the predictors through updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, the proposed procedure is expected to perform better than the marginal screening procedures when there are some predictors that are marginally independent of the survival time, but not jointly independent of the survival time. Meanwhile, since each iteration in Step 2 can avoid large-scale matrix inversion and, therefore, it can be carried out with low computational costs. Based on our simulation study, the proposed procedures can be implemented with less computing time than the marginal screening procedure studied in Fan, Feng and Wu (2000) and Zhao and Li (2012) in some scenarios (see Tables 3.2 and 3.3 for details). Theorem 1 below offers convergence behavior of the proposed algorithm.

Theorem 1. *Suppose that Conditions (D1)—(D4) in the Appendix hold. Denote*

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \left[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell_p''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\} \right]$$

where $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a matrix A. If $u_t \geq \rho^{(t)}$, then

$$\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}^{(t)}),$$

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e. $\|\beta\|_0 \leq m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem also provides us some insights about choosing u_t in practical implementation. In our numerical studies, this algorithm typically converges within six iterations. It is worth noting that Theorem 1 does not implies that the proposed algorithm converges to converge to the global optimizer.

2.2 Sure screening property

For the convenience of presentation, we use s to denote an arbitrary subset of $\{1, \ldots, p\}$, which amounts to a submodel with covariates $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. Also, we use $\tau(s)$ to indicate the size of model s. In particular, we denote the true model by $s^* = \{j : \beta_j^* \neq 0, 1 \leq j \leq p_n\}$ with $\tau(s^*) = \|\boldsymbol{\beta}^*\|_0 = q$. The objective of feature selection is to obtain a subset $\widehat{s}$ such that $s^* \subset \widehat{s}$ with a very high probability.