Published in final edited form as:

J Nonparametr Stat. 2010 ; 22(4): 499–515. doi:10.1080/10485250903348781.

Penalized variable selection with U-estimates

Xiao Song^{a,*} and Shuangge Ma^b

^a Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA

^b School of Public Health, Yale University, New Haven, CT, USA

Abstract

U-estimates are defined as maximizers of objective functions that are U-statistics. As an alternative to M-estimates, U-estimates have been extensively used in linear regression, classification, survival analysis, and many other areas. They may rely on weaker data and model assumptions and be preferred over alternatives. In this article, we investigate penalized variable selection with U-estimates. We propose smooth approximations of the objective functions, which can greatly reduce computational cost without affecting asymptotic properties. We study penalized variable selection using penalties that have been well investigated with M-estimates, including the LASSO, adaptive LASSO, and bridge, and establish their asymptotic properties. Generically applicable computational algorithms are described. Performance of the penalized U-estimates is assessed using numerical studies.

Keywords

U-estimate; penalization; variable selection

1. Introduction

With many statistical models, the unknown regression coefficient β can be estimated with $\hat{\beta} = \arg\max H_n(\beta)$, where the objective function H_n is a U-statistic computed based on n iid observations. $\hat{\beta}$ so obtained has been referred to as a *U-estimate*. In what follows, we provide representative examples of U-estimates.

Example 1

Consider the linear regression model $Y = \beta^T \mathbf{Z} + \varepsilon$, where β^T is the transpose of the length d regression coefficient β , $\mathbf{Z}$ is the length d vector of covariate, and ε is the random error. Assume n iid observations $(\mathbf{Z}_i, Y_i)$, $i = 1 \dots n$. Commonly adopted objective functions include $-\sum_i (Y_i - \beta^T \mathbf{Z}_i)^2$ and $-\sum_i |Y_i - \beta^T \mathbf{Z}_i|$, and the maximizers of these objective functions minimize the mean square and median of errors, respectively. Another objective function, which can be preferred under certain scenarios, is

$$H_n(\beta) = \frac{1}{n^2} \sum_{i,j} I(Y_i \geq Y_j) I(\beta^T \mathbf{Z}_i \geq \beta^T \mathbf{Z}_j),$$

*Corresponding author. xsong@uga.edu.

where I is the indicator function. Here H_n measures the *rank correlation* between Y and $\beta^T \mathbf{Z}$, and the estimate obtained by maximizing H_n is a special case of the maximum rank correlation (MRC) estimates defined in (1).

Example 2

Consider binary classification, where $Y \in \{0, 1\}$, $Pr(Y = 1|\mathbf{Z}) = G(\beta^T \mathbf{Z})$, and G is the monotone increasing link function. When G is known (e.g. the logit function), likelihood approaches can be used for estimation. However, parametric models and likelihood approaches may suffer from model misspecification. A more robust approach is to estimate β by maximizing

$$H_n(\beta) = \sum_{i,j} I(Y_i > Y_j) I(\beta^T \mathbf{Z}_i > \beta^T \mathbf{Z}_j) / \sum_i I(Y_i = 1) \sum_i I(Y_i = 0).$$

The objective function defined above can also be derived under the MRC framework. In addition, interestingly enough, $H_n(\beta)$ is also the *empirical AUC* (area under the curve) under the ROC (receiver operating characteristic) framework (3). As a robust alternative to parametric classification models, the ROC curve and corresponding empirical AUC have drawn extensive attentions in recent years.

Example 3

Consider a survival study, where T and C denote the event time of interest and censoring time, respectively. Assume the accelerated failure time (AFT) model, where $\log(T) = \beta^T \mathbf{Z} + \varepsilon$. Under case I interval censoring, one observation consists of $(C, I(T \leq C), \mathbf{Z})$. Such data are also known as the current status data. The rank estimate proposed by Aragon and Quiroz (4) and Abrevaya (5) maximizes

$$H_n(\beta) = \sum_{i,j} I(T_i \leq C_i) I(\log(C_i) - \beta^T \mathbf{Z}_i \geq \log(C_j) - \beta^T \mathbf{Z}_j) + I(T_j \leq C_j) I(\log(C_j) - \beta^T \mathbf{Z}_j \geq \log(C_i) - \beta^T \mathbf{Z}_i).$$

When the distribution of ε is known, the rank approach is valid, but may be inferior to likelihood approaches. However, when the distribution of ε is unknown, unlike likelihood approaches, the rank approach is still applicable.

Example 4

Consider a survival study, where T and C denote the event time and censoring time, respectively. Under the nonparametric transformation model, $g(T) = \beta^T \mathbf{Z} + \varepsilon$, where g is the *unknown* monotone transformation function. Under right censoring, one observation consists of $(V = \min(T, C), \Delta = I(T \leq C), \mathbf{Z})$. The rank estimate proposed by Khan and Tamer (2007) maximizes

$$H_n(\beta) = \sum_{i \neq j} I(T_i \leq C_i) I(V_i \geq V_j) I(\beta^T \mathbf{Z}_i \geq \beta^T \mathbf{Z}_j). \quad (1)$$

This estimate is especially desirable when the distribution of ε is unknown.

Examples 1–4 cover linear regression, binary classification, and survival analysis. More examples can be found in the literature. It has been proved that, many U-estimates, including those in Examples 1–4, have the much desired $\sqrt{n}$ consistency and asymptotic normality properties [1–2].

Literature review suggests that existing U-estimate studies have been focusing on estimation and inference. In practical data analysis, it is common to have multiple covariates, where only a subset of them are associated with the response of interest. With such data, variable selection is needed to discriminate important covariates from unimportant ones. For M-estimates, penalized selection approaches have drawn extensive attentions in recent years.

In this article, we investigate penalized variable selection when the objective function is a U-statistic. The penalized estimate is defined as

$$\widehat{\beta} = \operatorname{argmax}\{H_n(\beta) - p_\lambda(\beta)\}. \quad (2)$$

Here, H_n is a U-statistic and p_λ is the penalty function that depends on the tuning parameter λ . With properly chosen tunings, certain components of $\widehat{\beta}$ can be exactly zero. Only covariates with nonzero estimated coefficients enter the final models. Variable selection can thus be achieved.

Under certain scenarios, U-estimates can be preferred over alternatives, or even be the only choice. When there are multiple covariates, variable selection can be of critical interest. U-estimates and penalized variable selection have been *separately* investigated in published papers. However, to our best knowledge, penalized variable selection with U-estimates has not been studied. Our study will show that many “familiar results”, which have been established for penalized variable selection with M-estimates, hold for U-estimates. Although the main results in this article are not “surprising”, they are reassuring in that they establish the validity of adopting penalized variable selection techniques in U-estimation.

The rest of the article is organized as follows. We describe the smoothly approximated U-estimates in Section 2. Penalized estimates are defined in Section 3. Asymptotic estimation and variable selection properties are established. Computational algorithms are described in Section 4. Numerical studies, including simulations and analysis of two datasets, are provided in Section 5. We conclude with discussion in Section 6. Proofs are provided in the Appendix.

2. Smoothly approximated U-Estimates

The objective functions defined in Examples 1–4 all include the term $I(\beta^T \mathbf{Z}_i \geq \beta^T \mathbf{Z}_j)$ (the format is slightly different in Example 3). Literature review suggests that such a term is common to U-estimates. When there are multiple covariates, direct maximization of the objective functions – which are not continuous – can be difficult. Brutal search, stochastic annealing, or the approach developed in (5) all have a high computational cost when there are a moderate to large number of covariates. To tackle the computational problem, we approximate the indicator functions that involve β with $s_n(\cdot)$, where

$$s_n(u) = s(u/\sigma_n) \text{ and } s(u) = 1/\{1 + \exp(-u)\}.$$

Here σ_n is the positive tuning parameter. With slight abuse of notation, we will still use H_n to denote the objective functions with the indicator functions involving β replaced by s_n . We note that, in theory, s can be any smooth, monotone function satisfying $\lim_{u \rightarrow -\infty} s(u) = 0$ and $\lim_{u \rightarrow +\infty} s(u) = 1$. In our study, we choose the above specific form of s because of its simplicity.

The sigmoid approximation makes the objective functions continuously differentiable. Many existing numerical methods, such as the Newton-Raphson and gradient search, are thus applicable for optimization. Intuitively, when σ_n is small, $s_n(\cdot)$ mimics the behaviors of $I(\cdot)$. We will show in Section 3 that $\sigma_n = o(n^{-1/2})$ guarantees a satisfactory approximation. Smooth approximations to U-estimations have been separately proposed for Examples 1, 2 and 4 in [12–14]. We propose using such an approximation in the context of penalized variable selection.

The U-estimates we consider are location and scale invariant. To avoid the identifiability problem and without loss of generality, the first element of β is restricted to 1, i.e., $\beta = (1, \theta^T)^T$. This indicates a positive correlation between the response and the first covariate (the covariate can be recoded if necessary). We call the first covariate the “anchor”. Our interest lies in estimation and inference of θ .

Having an anchor requires at least one covariate to be associated with the response and the ability to identify it. In most if not all studies of U-estimates, it has been assumed that those two requirements can be satisfied. Theoretically, the first requirement can be checked with the hypothesis testing of $H_0: \|\beta\| = 0$. However, to our knowledge, a testing approach that is generically applicable has not been well developed. In practical data analysis, the first requirement can be usually satisfied. In Ma and Huang (2007), it has been proposed that the anchor can be selected by fitting marginal models for each covariate, and the covariate with the highest marginal significance will be selected as anchor. Ma and Huang (2007) suggests reasonable performance of this approach. The occasional inconvenience of fixing the anchor is the cost of adopting U-estimates, which usually demand assumptions much weaker than estimates (e.g. M-estimates) that do not need the anchor.

3. Penalized variable selection

Assume there are multiple covariates. We are interested in estimating θ , when some components of θ are zero and correspond to covariates that are not associated with the response. Without loss of generality, write $\theta = (\theta^1, \theta^2)^T$, where θ^1 contains the d_1 nonzero elements, θ^2 contains the d_2 zero elements, and $d_1 + d_2 = d$. Note that, with slight abuse of notation, we also use d to denote the length of θ . With penalized variable selection, we maximize

$$Q_n(\beta) = H_n(\beta) - \sum_{k=1}^d p_\lambda(|\theta_k|). \quad (3)$$

Popular penalty functions include:

1. LASSO. $p_\lambda(\theta) = \lambda|\theta|$. The LASSO is first proposed by Tibshirani (1996), and has been extensively investigated since then. For M-estimations, Knight and Fu (2004) establish that the LASSO is capable of variable selection. Leng et al. (2007) show that the LASSO is in general not variable selection consistent, unless stronger data and model assumptions are made.

2. Adaptive LASSO. $p_\lambda(\theta_k) = \lambda|\tilde{\theta}_k| - \gamma\theta_k$, where $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_d)$ is a $\sqrt{n}$ consistent estimator of θ (e.g. the ordinary U-estimate), and $\gamma > 0$. The adaptive LASSO is proposed to remedy the inconsistent selection problem of the LASSO. Zou (9) shows that, if a proper “starting value” $\tilde{\theta}$ is used, the adaptive LASSO is selection consistent, even when the LASSO is not.
3. Bridge. $p_\lambda(\theta) = \lambda\theta^\gamma$, $0 < \gamma < 1$. The Bridge penalty is also investigated in (6). In a recent study, with linear regression models and M-estimates, Huang et al. (10) establish the variable selection and estimation consistency.

Beyond the references listed above, the LASSO, adaptive LASSO, and bridge have been the focus of many other studies. However, we note that, published literature has been focusing on M-estimates and little attention has been paid to U-estimates.

Of note, there are other penalty functions that have been extensively used, for example the SCAD. In this article, we focus on the above three type of penalties, because of their popularity and “bridge”-like format. We expect that investigations of other penalties can be conducted in a similar manner.

All the three penalties we investigate have been proposed in the context of M-estimation. A question of critical interest is whether new penalties are needed with U-estimates. The essential idea of the above three penalties (and several others) is to control the complexity of the models by controlling the norms of regression coefficients. Since both M- and U-estimations are just tools to recover the original models, it seems to us that the same penalty functions can be used under different estimation methods. Our preliminary investigation does not suggest any new penalties, although such a possibility may exist.

4. Asymptotic properties

U-estimates represent a large family of estimates applied to a great variety of statistical models. Although there exist some results that are applicable to many different U-estimates, we closely examine Examples 1–4, but fail to construct a unified set of data and model assumptions and proofs. Thus, we take a rather laborious approach and investigate the four examples separately. We find that results presented in Theorems 4.1 and 4.2 hold for all examples we investigate. In addition, the assumptions on the tuning parameters (σ_n and λ) are the same across all examples. However, certain technical details, for example forms of asymptotic variances and parts of the proofs, differ significantly. In what follows, we provide detailed results and proofs for Example 4, which has the most complicated statistical model among the four, *with the understanding that other examples can be investigated in a similar manner with only minor modifications.*

The estimate defined in (1) has been referred to as the partial rank (PR) estimate in (16). Accordingly, we refer to the estimate defined in (3) as the penalized smoothed partial rank (PSPR) estimate. To indicate that θ is the actual parameter, we write $\beta(\theta) = (1, \theta^T)^T$. Let

$\beta_0 = (1, \theta_0^T)^T$ denote the true value of β . Let $X = (V, \Delta, \mathbf{Z})$ and $x = (v, \delta, \mathbf{z})$. Let R be the support operator, for example, $R(\mathbf{Z})$ denotes the support of $\mathbf{Z}$. Let $\mathbf{Z}_{(-1)}$ denote the last d components of covariate vector $\mathbf{Z}$, let $g^0(w|r)$ denote the conditional density of $W = \beta_0^T \mathbf{Z}$ given $\mathbf{Z}_{(-1)} = r$, and let $p^0(\delta, y, w|r)$ denote the conditional density of (Δ, Y, W) given $\mathbf{Z}_{(-1)} = r$. Define

$$\begin{aligned}\tau(x, \beta(\theta)) &= E\{\Delta I(V \geq V) I(\beta^T \mathbf{z} - \beta^T \mathbf{Z} > 0) + \delta I(V \geq v) I(\beta^T \mathbf{Z} - \beta^T \mathbf{z} > 0)\}, \\ H(\beta) &= E\{\Delta_j I(V_i \geq V_j) I(\beta^T \mathbf{Z}_i > \beta^T \mathbf{Z}_j)\}.\end{aligned}$$

We assume:

- A1** The set $\{\mathbf{Z} \in R(\mathbf{Z}): Pr(\Delta = 1|\mathbf{Z}) > 0\}$ has a positive measure.
- A2** The random error ε is independent of C and $\mathbf{Z}$.
- A3** The first component of $\mathbf{Z}$ has an everywhere positive Lebesgue density, conditional on other components.
- A4** The parameter space $\mathbb{B}$ containing β_0 is a compact subset of $\mathbb{R}^{d+1}$.
- A5** $R(\mathbf{Z})$ is not contained in any proper linear subspace of $\mathbb{R}^{d+1}$.
- A6** T and C are conditionally independent given $\mathbf{Z}$.
- A7** (a) For each x , the function $\tau(x, \beta(\theta))$ is twice differentiable with respect to θ in a neighborhood of θ_0 with the k^{th} derivative $\nabla_k \tau(x, \beta(\theta))$, $k = 1, 2$. The second order derivative $\nabla_2 \tau(x, \beta(\theta))$ satisfies the Lipschitz condition. (b) The partial derivatives of $g^0(w|r)$ and $p^0(\delta, y, w|r)$ with respect to w exist and are bounded.
- A8** $E\|\nabla_1 \tau(x, \beta(\theta_0))\|^2$ and $E\|\nabla_2 \tau(x, \beta(\theta_0))\|$ are finite, and $E\{\nabla_2 \tau(x, \beta(\theta_0))\}$ is non-singular.
- A9** $\sigma_n = o(n^{-1/2})$.
- B1** The penalty function satisfies that $p_\lambda(0) = 0$, $p_\lambda(\theta) \geq 0$ and $p_\lambda(\theta)$ is twice differentiable for $\theta > 0$.

A1–A9 are assumed in (14) to establish the asymptotic normality of the smoothed PR estimator. A1–A6 are standard for nonparametric transformation models (Khan and Tamer, 2007). For practical data, A1, A2, A3 and A5 can be partly checked with the design of the study. If the parameter set being considered is large enough, A4 can be satisfied. A7 and A8 are mainly made to guarantee non-singularity of the asymptotic variance of the estimate. A9 limits the convergence rate of the smoothing parameter. Numerical studies in (14) and others show that performance of the smooth approximation is not sensitive to the value of σ_n . In practice, we suggest $\sigma_n = a \times n^{-1/2}$ with a small positive a . B1 gives general restrictions on the penalty function. Among the above assumptions, A1, A2, A6 and A7 are specific to the nonparametric transformation model and right censored data, whereas others are more “generic”. For the U-estimates in Examples 1–3, comparable but significantly different assumptions need to be made.

Denote the first and second derivatives of $p_\lambda(\theta)$ by $p'_\lambda(\theta)$ and $p''_\lambda(\theta)$, respectively. Denote the j th component of θ_0 as θ_{j0} . Let

$$a_n = \max\{p'_\lambda(|\theta_{j0}|): \theta_{j0} \neq 0\} \text{ and } b_n = \max\{p''_\lambda(|\theta_{j0}|): \theta_{j0} \neq 0\}.$$

We first establish the consistency as follows.

Theorem 4.1

(Consistency) Suppose A1–A9 and B1 hold. If $b_n \rightarrow 0$, then there exists a local maximizer $\hat{\theta}$ of Q_n such that $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2} + a_n)$.

With the LASSO penalty, $a_n = \lambda$ and $b_n = 0$. With the adaptive LASSO penalty, $a_n = \lambda \max_{\theta_{j0} \neq 0} \{|\theta_{j0}|^{-\gamma}\}$ and $b_n = 0$. With the bridge penalty, $a_n = \lambda \gamma \max_{\theta_{j0} \neq 0} \{|\theta_{j0}|^{\gamma-1}\}$ and $b_n = \lambda \gamma (\gamma - 1) \min_{\theta_{j0} \neq 0} \{|\theta_{j0}|^{\gamma-2}\}$. Thus when $\lambda = o(1)$, the penalized U-estimates with all three

penalties are consistent. The $\sqrt{n}$ consistency further requires $\lambda = O(n^{-1/2})$. Built on Theorem 4.1, we can establish the variable selection property as follows.

Theorem 4.2

(Oracle property) Suppose

$$a_n = O(n^{-1/2}), \quad (4)$$

$$\sup_{0 < \|\theta_k\| = O(n^{-1/2})} \{n^{1/2} p'_\lambda(\theta_k)\}^{-1} = o_p(1) \text{ for } k = d_1 + 1, \dots, d. \quad (5)$$

Then under assumptions of Theorem 4.1, with probability going to 1, the $\sqrt{n}$ consistent estimator $\hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2)$ in Theorem 4.1 satisfies

- (Sparsity) $\hat{\theta}^2 = 0$;
- (Asymptotic normality) $n^{-1/2}\{A - 2D_2(\theta_0^1)\} \{\hat{\theta}^1 - \theta_0^1\} + D_1(\theta_0^1) \rightarrow N(0, B)$, where

$$\begin{aligned} A &= -E\{\nabla_2 \tau(x, \beta(\theta_0))\}/2, \\ B &= E\{\nabla_1 \tau(x, \beta(\theta_0)) \nabla_1^T \tau(x, \beta(\theta_0))\} \\ D_1(\theta_0^1) &= \text{diag}\{p'_\lambda(|\theta_{10}|) \text{sgn}(\theta_{10}), \dots, p'_\lambda(|\theta_{d_{10}}|) \text{sgn}(\theta_{d_{10}})\}, \\ D_2(\theta_0^1) &= \text{diag}\{p''_\lambda(|\theta_{10}|), \dots, p''_\lambda(|\theta_{d_{10}}|)\}. \end{aligned}$$

Theorem 4.2 provides sufficient conditions for the oracle variable selection property, and can be used to establish such property of the adaptive LASSO and bridge. More specifically, with the adaptive LASSO penalty,

$$\sup_{0 < \|\theta_k\| = O(n^{-1/2})} \{n^{1/2} p'_\lambda(\theta_k)\}^{-1} = \max_{\theta_k \neq 0} (n^{1/2} \lambda |\tilde{\theta}_k|^{-\gamma})^{-1}.$$

Note that $\tilde{\theta}_k = O_p(n^{-1/2})$ for $k = d_1 + 1, \dots, d$. Thus when $\lambda = O(n^{-1/2})$, conditions (4) and (5) hold. With the bridge penalty,

$$\sup_{0 < \|\theta_k\| = O(n^{-1/2})} \{n^{1/2} p'_\lambda(\theta_k)\}^{-1} = \sup_{0 < \|\theta_k\| = O(n^{-1/2})} \{n^{1/2} \lambda \gamma \theta_k^{\gamma-1}\}^{-1} \leq \{\lambda \gamma O(n^{1-\gamma/2})\}^{-1}.$$

So when $\lambda = O(n^{-1/2})$, conditions (4) and (5) hold.

With the LASSO penalty, $\sup_{0 < \|\theta_k\| = O(n^{-1/2})} \{n^{1/2} p'_\lambda(\theta_k)\}^{-1} = (n^{1/2} \lambda)^{-1}$. However, condition (4) requires that $\lambda = O(n^{-1/2})$, under which condition (5) does not hold. Thus, Theorem 4.2 does not imply the oracle property of the LASSO. Following Leng et al. (8), it can be proved that the LASSO does not have the oracle property, unless stronger model assumptions are made. Such results are of less interest and will not be further pursued.

As sample size goes to infinity, $Q_n(\beta)$ converges to its expectation, which has a unique maximizer. When the sample size is large enough, we expect a unique maximizer for $Q_n(\beta)$. However, when the sample size is extremely small, under rare scenarios, there may be multiple maximizers for $Q_n(\beta)$. We state Theorems 4.1 and 4.1 in terms of “local maximizers” to emphasize that, when there exist multiple maximizers, we are only interested in the one in the neighborhood of the true parameter. The localization statement is mainly for theoretical rigorousness. In our numerical studies, when the sample size is reasonable, multiple maximizers have not been observed.

5. Computational algorithms

5.1. LASSO

The LASSO penalized U-estimate defined in (3) is equivalent to

$$\widehat{\beta} = \operatorname{argmax}_{\beta} H_n(\beta) \text{ with } \sum_k |\beta_k| \leq M,$$

where there is a one-to-one correspondence between M and λ . Under this correspondence, the LASSO penalized U-estimate can be computed using the L_1 boosting algorithm proposed by Kim and Kim (17). Write $H_n(\beta) = H_n(\beta(\theta), \mathbf{Z})$.

1. Initialize $\theta = 0$ and $s = 0$.
2. With the current estimate of θ , compute $\varphi(\theta, \mathbf{Z}) = \partial H_n(\beta(\theta), \mathbf{Z}) / \partial \theta$. Denote the k^{th} component of φ as φ^k .
3. Find k^* that maximizes $|\varphi^k(\theta, \mathbf{Z})|$. If $\varphi^{k^*}(\theta, \mathbf{Z}) = 0$, stop the iteration.
4. Denote $\gamma = \operatorname{sign}(\varphi^{k^*}(\theta, \mathbf{Z}))$. Find $\hat{\pi}$ such that

$$\hat{\pi} = \operatorname{argmax}_{\pi \in [0,1]} H_n((1 - \pi)\theta + \pi M \gamma \eta^{k^*}),$$

where η^{k^*} is a length d vector with the k^{*th} element equal to 1 and the rest equal to 0.

5. Let $\theta_k = (1 - \hat{\pi})\theta_k$ for $k \neq k^*$ and $\theta_{k^*} = (1 - \hat{\pi})\theta_{k^*} + \gamma M \hat{\pi}$. Let $s = s + 1$.
6. Repeat steps 2–5 until convergence.

The θ at convergence is the LASSO estimate. We conclude convergence if the absolute value of $\varphi^{k^*}(\beta)$ computed in step 3 is less than a pre-defined criteria, and/or if the absolute difference of θ between two consecutive iterations is smaller than a pre-defined threshold. For M-estimates, there are computational algorithms (e.g. LARS, coordinate descent) more efficient than the boosting. One advantage of the boosting algorithm is its generic applicability. In addition, it is relatively insensitive to the number of covariates, and hence preferred in studies with multiple covariates.

5.2. Adaptive LASSO

The adaptive LASSO estimator can be obtained using the same algorithm as for LASSO via a transformation of the parameters. Let Z_k be the k^{th} component of $\mathbf{Z}$.

1. Replace Z_{k+1} by $Z_{k+1}^\# = |\tilde{\theta}_k|^\gamma Z_{k+1}$, $k = 1, \dots, d$.

2. Apply LASSO with $\mathbf{Z}=(Z_1, Z_2^{\#}, \dots, Z_{d+1}^{\#})^T$ as the covariates. Denote the θ at convergence as $\widehat{\theta}^{\#}=(\widehat{\theta}_1^{\#}, \dots, \widehat{\theta}_d^{\#})$.
3. The adaptive LASSO estimator is $(|\widehat{\theta}_1|^\gamma \widehat{\theta}_1^{\#}, \dots, |\widehat{\theta}_d|^\gamma \widehat{\theta}_d^{\#})^T$.

5.3. Bridge

The bridge estimator can be obtained by modifying the algorithm for the LASSO estimator.

Note that the bridge penalty function $\lambda \sum |\theta_j|^\gamma \approx \lambda \sum \frac{|\theta_j|}{|\widehat{\theta}_j|^{1-\gamma}}$, if $\widehat{\theta}_j$ is close to θ_j . This motivates the following algorithm.

1. Initialize $\theta = 0$, $\widehat{\theta}_j = \widetilde{\theta}_j$, and $s = 0$. Here $\widetilde{\theta}_j$ is the initial $\sqrt{n}$ consistent estimate of θ_{j0} and can be set as the ordinary U-estimate.
2. Let $Z_{k+1}^* = |\widehat{\theta}_k|^{1-\gamma} Z_{k+1}$ ($k=1, \dots, d$) and $\mathbf{Z}^*=(Z_1, Z_2^*, \dots, Z_{d+1}^*)$. Compute $\varphi(\theta, \mathbf{Z}^*) = \partial H_n(\beta(\theta), \mathbf{Z}^*)/\partial \theta$. Denote the k^{th} component of φ as φ^k .
3. Find k^* that maximizes $|\varphi^k(\theta, \mathbf{Z}^*)|$. If $\varphi^{k^*}(\theta, \mathbf{Z}^*) = 0$, stop the iteration.
4. Denote $\gamma = \text{sign}(\varphi^{k^*}(\theta, \mathbf{Z}^*))$. Find $\widehat{\pi}$ such that

$$\widehat{\pi} = \arg\max_{\pi \in [0,1]} H_n((1-\pi)\theta + \pi M \gamma \eta^{k^*}),$$

where η^{k^*} is a length d vector with the k^{*th} element equal to 1 and the rest equal to 0.

5. Let $\theta_k = (1 - \widehat{\pi})\theta_k$ for $k \neq k^*$ and $\theta_{k^*} = (1 - \widehat{\pi})\theta_{k^*} + \gamma M \widehat{\pi}$. Let $\widehat{\theta}_k = \theta_k I(|\theta_k| > \varepsilon) + \varepsilon I(|\theta_k| \leq \varepsilon)$ for some small $\varepsilon > 0$, $k = 1, \dots, d$. Let $s = s + 1$.
6. Repeat steps 2–5 until convergence or a fixed number of iterations has been reached. Denote the θ at convergence as $\widehat{\theta}^{\#}=(\widehat{\theta}_1^{\#}, \dots, \widehat{\theta}_d^{\#})$.
7. The bridge estimator is $(|\widehat{\theta}_1|^{1-\gamma} \widehat{\theta}_1^{\#}, \dots, |\widehat{\theta}_d|^{1-\gamma} \widehat{\theta}_d^{\#})^T$.

6. Numerical studies

6.1. Simulation

We conduct simulation studies to assess performance of the penalized variable selection with U-estimates. As in Section 3, we focus on Example 4. Comparable performance is expected for other examples. Penalized estimation involves the tuning parameter λ (or M). In the numerical study, we use the K-fold cross validation for tuning parameter selection. Our numerical study suggests its satisfactory performance. Since it is not the focus of our study, we will not investigate alternative tuning parameter selection methods and the asymptotic properties. We consider the case of seven covariates $Z_1, \dots, Z_7$, which have independent standard normal distributions. We consider the following three sets of simulations.

Set I—The survival time T depends on the covariates through a Cox model with the regression coefficients $(-1, -0.8, -0.6, 0, 0, 0, 0)$ and the baseline hazard equal to $\exp(0.5t)$. This corresponds to the nonparametric transformation model with $\theta = (0.8, 0.6, 0, 0, 0, 0, 0)$, which includes $d_1 = 2$ nonzero elements and $d_2 = 4$ zero elements. The censoring time C is generated from an exponential distribution with mean 4, leading to a censoring rate of

19.3%. We set the smoothing parameter $\sigma = n^{-1/2}/3$, $\gamma = 1$ for the adaptive LASSO estimator, and $\gamma = 1/2$ for the bridge estimator.

Set II—We consider an AFT model with $\theta = (0.8, 0.6, 0, 0, 0, 0)$, where the error is generated from a normal distribution with mean 0 and standard deviation 0.1. The censoring time C is generated from an exponential distribution with mean 4, leading to a censoring rate of 31.4%.

Set III—Set III differs from Set II only in the transformation function $g(t)$. We set $g(t) = \sqrt{t} - 2$. The censoring rate is 56.9%. This model is neither a Cox model nor an AFT model.

Under each set of simulation, we generate 200 datasets with sample size equal to 100. The tuning parameters are selected using the 3-fold cross validation. For comparison, we also fit the Cox model and the AFT model with the log transformation. Of note, under Set I, the AFT model is mis-specified; under Set II, the Cox model is mis-specified; and under Set III, both the Cox and AFT models are mis-specified.

We summarize the variable selection results in Table 1. We can see that (a) as expected, the Cox, AFT, and ordinary U-estimation are not capable of variable selection; (b) the LASSO can identify all covariates with nonzero coefficients. In addition, it also has nonzero false positive rates, which can be as high as $\sim 40\%$; (c) the adaptive LASSO performs as well as the LASSO in identifying the nonzero coefficients with a rate of 100%, and much better in correctly identifying the zero coefficients; and (d) the bridge performs even better in identifying the zero coefficients than the adaptive LASSO. However, for the third covariate, since the regression coefficient is small (0.6), there is a small probability of mistakenly identifying it as zero. Estimation results are summarized in Table 2. We can see that all penalized approaches yield reasonable estimation results.

Remark—Simulations suggest that all three penalties can lead to reasonable variable selection performance with U-estimates. The adaptive LASSO and bridge penalties outperform the LASSO by selecting a smaller number of false positives. Of note, such an observation is similar to that with M-estimates. In addition, estimation results in Table 2 further demonstrate the merits of U-estimates, which have been established in published literature. For example in simulation Set II, the underlying model is the AFT model. Fitting the Cox model, which is the default in many studies, can lead to highly biased estimates. The U-estimate relies on weaker assumptions and is consistent.

6.2. Analysis of PBC data

Consider the primary biliary cirrhosis of the liver (PBC) study conducted by Mayo Clinic between January 1974 and May 1984 (18). The objective was to compare the drug D-penicillamine (DPCA) with a placebo on the time to death for PBC patients. Information on sixteen clinical, bio-mechanical, serologic and histologic variables are collected on 312 patients. Among the 312 patients, 276 have complete observations on the survival and covariates $Z_1 = \text{age}$, $Z_2 = \text{serum albumin}$, $Z_3 = \text{alkaline phosphatase}$, $Z_4 = I(\text{presence of ascites})$, $Z_5 = \text{serum bilirubin}$, $Z_6 = \text{serum cholesterol}$, $Z_7 = \text{presence of edema}$, $Z_8 = I(\text{enlarged liver})$, $Z_9 = \text{platelet count}$, $Z_{10} = \text{standardized blood clotting time}$, $Z_{11} = I(\text{sex} = \text{male})$, $Z_{12} = \text{liver enzyme}$, $Z_{13} = I(\text{blood vessel malformations in the skin})$, $Z_{14} = I(\text{treatment} = \text{DPCA})$, $Z_{15} = \text{triglycerides}$, $Z_{16} = \text{urine copper}$. Fleming and Harrington (18) shows that only a subset of those covariates are associated with survival. Variable selection is warranted.

We assume the nonparametric transformation model, adopt the U-estimate, and use the LASSO, adaptive LASSO, and bridge penalties for variable selection. For comparison, we also fit the Cox model, AFT model, and “transformation model + smoothed U-estimate (without penalty)”. We use $\gamma = 1$ for the adaptive LASSO and $\gamma = 1/2$ for the bridge. We normalize the continuous covariates to have mean 0 and standard deviation 1. The estimates are shown in Table 3. The LASSO, adaptive LASSO, and bridge penalized U-estimates have 14, 11, and 3 nonzero components, respectively.

To assess performance of the penalized variable selection, we randomly split data into a training set and a testing set with equal sizes. With each penalty, we compute the penalized estimate using the training set. We then make predictions for subjects in the testing set. We dichotomize $\beta^T Z$ at the median, create two risk sets, and evaluate the difference of survival functions of the two sets using the logrank statistic. By doing so, we can assess whether the estimates have predictive power. To avoid bias caused by an extreme split, this process is repeated for 100 randomly split training and testing set pairs. The medians of the 100 logrank statistics are 23.14, 25.09, 24.22, 25.53, 25.67, 24.07, respectively, for the Cox, AFT, smoothed U-estimate, LASSO, adaptive LASSO, and bridge, respectively. We can conclude that the models obtained using penalized U-estimations have similar prediction performance as those under Cox, AFT and U-estimation, while using a much smaller number of covariates.

6.3. Analysis of UIS data

We consider the UIS data described in (19), which is a subset of data from the University of Massachusetts Aids Research Unit IMPACT Study. The study involves two concurrent randomized trials on drug users at two sites A and B with sample sizes 444 and 184, respectively. Each site focuses on a separate pair of treatments with short or long durations. The objective is to compare the treatments on their effects to reduce the time to return to drug use. Information is also collected on age, race, Beck Depression score, heroin/cocaine use during three months prior to admission (hercoc), IV drug use at admission (ivhx), prior drug treatment (table 1.4, (19)). A total of 12 covariates are considered: Z_1 = length of treatment, Z_2 = age, Z_3 = Beck Depression score, Z_4 = number of prior drug treatments, $Z_5 = I(\text{race} = \text{white})$, $Z_6 = I(\text{treatment} = \text{short})$, $Z_7 = I(\text{site} = \text{B})$, $Z_8 = I(\text{hercoc} = \text{heroin and cocaine})$, $Z_9 = I(\text{hercoc} = \text{heroin only})$, $Z_{10} = I(\text{hercoc} = \text{covaine only})$, $Z_{11} = I(\text{ivhx} = \text{never})$, and $Z_{12} = I(\text{ivhx} = \text{previous})$.

For the 575 subjects with complete observations, we fit the same set of models as for the PBC data. The continuous covariates are normalized to have mean 0 and standard deviation 1. The estimates are shown in Table 4. The LASSO, adaptive LASSO, and bridge penalties select 10, 8, and 3 covariates, respectively. The logrank statistics calculated based on 100 randomly split training and testing set pairs have medians of 55.52, 53.87 54.09, 52.81, 51.46, 50.38 for the Cox, AFT, smoothed U-estimate, LASSO, adaptive LASSO, and bridge, respectively.

7. Discussion

U-estimation has been applied to a great variety of statistical models. U-estimates may rely on weaker assumptions and be preferred under certain scenarios. In this article, we investigate penalized variable selection with U-estimates. For a class of U-estimates that share similar formats as those in Examples 1–4, the objective functions are not continuous. We propose smooth approximations to the objective functions, which makes them continuously differentiable. We then consider using the LASSO, adaptive LASSO, and bridge for penalized variable selection. In this article, it is not our intention to “reinvent” the penalized variable selection approaches. Rather, we intend to show that, penalized

approaches that have been developed with M-estimates are applicable to U-estimates. Considering the significant difference between M-estimates and U-estimates, this study is warranted beyond the extensive studies on penalized M-estimates.

We show that for U-estimates in Examples 1–4, the adaptive LASSO and bridge penalized approaches have the oracle variable selection property. This article is among the first to establish that the variable selection properties of penalized approaches with U-estimates are comparable to their counterparts with M-estimates. The similarity (between results for M-estimates and U-estimates) is reassuring, and suggests that many well developed penalized selection techniques may be transferred from M-estimates to U-estimates. To avoid confusion of notations, we provide detailed proofs for Example 4 only. Other examples can be studied in a similar manner. For M-estimates, it has been shown that properties of variable selection approaches need to be established on a case-by-case basis. Thus, we do not expect unified results for all U-estimates. Our numerical studies show that the LASSO has reasonable variable selection properties. The adaptive LASSO and bridge out-perform LASSO by identifying fewer false positives.

In this article, we focus on estimation. Inference is not investigated. For penalized M-estimators, the following strategies have been suggested for inference of $\hat{\theta}^1$: (a) a plug-in approach. Intuitively, the variance can be obtained by directly estimating the asymptotic variance with true parameters replaced with estimated values; (b) bootstrap, which has been the choice of inference for studies such as (20). For U-estimates without penalization, both the plug-in approach and the bootstrap have been suggested for inference (21). Inference for the penalized U-estimates can be fundamentally different from published articles because of the format of the objective functions and the penalty. To our best knowledge, there is no inference technique that has been theoretically justified for penalized U-estimates. The following possibilities have been considered: (a) a plug-in approach. Song et al. (14) show that such an approach even fails to provide satisfactory inference for the U-estimate, unless the sample size is extremely large; (b) bootstrap. Both the weighted and nonparametric bootstrap have been considered. Our numerical studies suggest that the above approaches fail to provide satisfactory empirical coverage (results not shown here). At this moment, it is not clear what inference method is valid under the present setup.

In this article, we assume that the number of covariates is finite. In some recent studies, the ultra-high dimensional case with d diverging with n has been considered. We suspect that, with a diverging number of covariates, the proposed penalized variable selection with U-estimates approach is still applicable. For penalized M-estimates, data assumptions and proofs of variable selection properties with a finite or a diverging number of covariates are significantly different. Investigation of the proposed methodology with $d \rightarrow \infty$ is highly nontrivial and will be pursued in future research.

Acknowledgments

The research is partially supported by the UGARF grant from University of Georgia (Song) and LM009754 from NIH (Ma). We would like to thank the editor, associate editor, and two referees for insightful comments, which have led to significant improvement of the paper.

References

1. Han AK. Non-parametric Analysis of a Generalized Regression Model. *J Econometrics*. 1987; 35:303–316.
2. Sherman RP. The limiting distribution of the maximum rank correlation estimator. *Econometrica*. 1993; 61:123–137.

3. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; 2003.
4. Aragon J, Quiroz AJ. Rank regression for current status data. *Statist Prob Letters*. 1995; 24:251–256.
5. Abrevaya J. Rank regression for current-status data: asymptotic normality. *Statist Prob Letters*. 1999; 43:275–287.
6. Knight K, Fu WJ. Asymptotics for LASSO-type estimators. *Ann Statist*. 2000; 28:1356–1378.
7. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001; 96:1348–1360.
8. Leng C, Lin Y, Wahba G. A note on the LASSO and related procedures in model selection. *Statist Sinica*. 2006; 16:1273–1284.
9. Zou H. The adaptive lasso and its oracle properties. *J Amer Statist Assoc*. 2006; 101:1418–1429.
10. Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann Statist*. 2008; 36:587–613.
11. Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high dimensional linear regression. *Ann Statist*. 2008; 36:1567–1594.
12. Horowitz JL. A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica*. 1992; 60:505–31.
13. Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics*. 2007; 63:751–757. [PubMed: 17825007]
14. Song X, Ma S, Huang J, Zhou XH. A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics*. 2007; 8:197–211. [PubMed: 16670240]
15. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Statist Soc Ser B*. 1996; 58:267–288.
16. Khan S, Tamer E. Partial rank estimation of duration models with general forms of censoring. *J Econometrics*. 2007; 136:251–280.
17. Kim, Y.; Kim, J. Gradient LASSO for feature selection. *Proceedings of the 21st International Conference on Machine Learning*; 2004.
18. Fleming, TR.; Harrington, DP. *Counting Processes and Survival Analysis*. Wiley; New York: 1991.
19. Hosmer, DW.; Lemeshow, S. *Applied Survival Analysis*. Wiley; New York: 1999.
20. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics*. 2006; 62:813–820. [PubMed: 16984324]
21. Arcones MA, Gine E. On the bootstrap of U and V statistics. *Ann Statist*. 1992; 20:655–674.
22. Koroljuk, VS.; Borovskich, YV. *Theory of U-Statistics*. Kluwer Academic Publishers; 1994.

Appendix A. Proof of Theorem 4.1

It has been shown in Song et al. (2007) that $H_n(\beta)$ converges uniformly to

$$H(\beta) = E\{\Delta_j I(V_i \geq V_j) I(\beta^T \mathbf{Z}_i > \beta^T \mathbf{Z}_j)\},$$

which has a unique maximizer β_0 . Under A7 and A8, $H(\beta(\theta))$ is twice differentiable around a neighborhood of θ_0 , $\nabla H(\beta(\theta_0)) = \mathbf{0}$ and $\nabla^2 H(\beta(\theta_0))$ is negative definite. It is sufficient to show that for any $\varepsilon > 0$, there exists a large constant M_1 such that

$$P \left\{ \sup_{\|u\|=M_1} Q_n(\beta_0 + h_n u) < Q_n(\beta_0) \right\} > 1 - \varepsilon,$$

where $h_n = n^{-1/2} + a_n$.

Since $p_\lambda(|\theta_k|) \geq p_\lambda(0) = 0$, we have

$$n^{-1} \{Q_n(\beta_0 + h_n u) - Q_n(\beta_0)\} \leq n^{-1} \{H_n(\beta_0 + h_n u) - H_n(\beta_0)\} - n^{-1} \sum_{\theta_{k0} \neq 0} \{p_\lambda(|\theta_{k0} + h_n u_k|) - p_\lambda(|\theta_{k0}|)\}. \quad (\text{A1})$$

Noting that

$$H_n(\beta) - H_n(\beta_0) \xrightarrow{a.s.} H(\beta) - H(\beta_0)$$

uniformly for β in a neighborhood of β_0 and $\nabla H(\beta(\theta_0)) = \mathbf{0}$, the first term in (A1) can be written as

$$H_n(\beta_0 + h_n u) - H_n(\beta_0) = h_n^2 u^T \{\nabla^2 H(\beta(\theta_0)) + o_p(1)\} u + o_p(1).$$

The second term is bounded by

$$\begin{aligned} & \left| \sum_{\theta_{k0} \neq 0} \{p_\lambda(|\theta_{k0} + h_n u_k|) - p_\lambda(|\theta_{k0}|)\} \right| \\ &= \left| \sum_{\beta_{j0} \neq 0} \{p'_\lambda(|\theta_{k0}|) h_n u_k \text{sgn}(\theta_{k0}) + p''_\lambda(|\theta_{k0}|) h_n^2 u_{j0}^2 \text{sgn}(\theta_{k0})\} + o_p(h_n^2) \right| \\ &\leq \left| \sum_{\beta_{j0} \neq 0} \{a_n h_n u_{j0} + b_n h_n^2 u_{j0}^2\} \right| + o_p(h_n^2) \\ &\leq d_1 \{a_n h_n \|u\| + b_n h_n^2 \|u\|^2\} + o_p(n^{-1} h_n^2) \\ &\leq d_1 h_n^2 \|u\| + o_p(h_n^2) \end{aligned}$$

as $b_n \rightarrow 0$. This implies that the second term in (A1) is dominated by the first term when M_1 is large. This establishes consistency of the estimator.

Appendix B. Proof of Theorem 4.2

We first show that the estimator has the sparsity property. From Theorem 4.1, we only need to show that

$$Q_n\{(\theta^{1T}, 0^{2T})^T\} = \max_{\|\theta^2\| \leq Cn^{-1/2}} Q_n\{(\theta^{1T}, \theta^{2T})^T\}$$

for $\theta^1 - \theta_0^1 = O_p(n^{-1/2})$. It is sufficient to show that $\partial Q_n / \partial \theta_k$ and θ_k have different signs for $\theta_k \in (-Cn^{-1/2}, Cn^{-1/2})$ for $k = d_1 + 1, \dots, d$. Note that for $\theta_k \neq 0$,

$$\frac{\partial Q_n}{\partial \theta_k} = \frac{\partial H_n(\beta(\theta))}{\partial \theta_k} - p'_\lambda(|\theta_k|) \text{sgn}(\theta_k).$$

Consider the first term. By a Taylor expansion,

$$\frac{\partial H_n(\beta(\theta))}{\partial \theta_k} = \frac{\partial H_n(\beta_0)}{\partial \theta_k} + \sum_{r=1}^d \frac{\partial^2 H_n(\beta^*)}{\partial \theta_k \partial \theta_r} (\theta_r - \theta_{r0}) \quad (\text{B1})$$

The first term in (B1) is

$$\frac{\partial H_n(\beta_0)}{\partial \theta_k} = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) s' \left\{ \beta_0^T \left(\frac{\mathbf{Z}_i - \mathbf{Z}_j}{\sigma_n} \right) \right\} \frac{\mathbf{Z}_{ik} - \mathbf{Z}_{jk}}{\sigma_n},$$

where $s'(u) = e^{-u}/(1 + e^{-u})^2$ is the first order derivative of $s(u)$. Note that we assume boundedness and the tuning parameter σ_n satisfying A9. $\frac{\partial H_n(\beta_0)}{\partial \theta_k}$ is a second order degenerate U-statistic. Following Koroljuk and Borovskich (22), we can prove that $\frac{\partial H_n(\beta_0)}{\partial \theta_k}$ converges with $\sqrt{n}$ rate. Hence

$$\frac{\partial H_n(\beta_0)}{\partial \theta_k} = O_p(n^{-1/2}).$$

Similarly,

$$\begin{aligned} \frac{\partial^2 H_n(\beta_0)}{\partial \theta_k \partial \theta_r} &= \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) s'' \left\{ \beta_0^T \left(\frac{\mathbf{Z}_i - \mathbf{Z}_j}{\sigma_n} \right) \right\} \frac{\mathbf{Z}_{ik} - \mathbf{Z}_{jk}}{\sigma_n} \frac{\mathbf{Z}_{ir} - \mathbf{Z}_{jr}}{\sigma_n} \\ &= O_p(n^{-1/2}), \end{aligned}$$

where $s''(u) = (e^{-2u} - e^{-u})/(1 + e^{-u})^3$ is the second order derivative of $s(u)$. Together with $\theta_r - \theta_{r0} = O_p(n^{-1/2})$, it follows that the second term in (B1) is $O_p(n^{-1/2})$. Hence we have

$$\frac{\partial H_n(\beta)}{\partial \theta_k} = O_p(n^{-1/2}) + O_p(n^{-1}) = O_p(n^{-1/2}).$$

Then

$$\frac{\partial Q_n}{\partial \theta_k} = O_p(n^{-1/2}) - p'_\lambda(\theta) \text{sgn}(\theta_k).$$

Under Condition (5), the second term is dominated. This completes the proof.

With the proof of the consistency, we have shown that there exists a $\hat{\beta}_1$ such that $\widehat{\beta}=(\widehat{\beta}_1^T, 0_{d_2}^T)^T$ maximizes $Q_n(\beta)$. From now on, we consider $\theta = (\theta_1, 0)$. Using the same notations as those in Section 4.1, we first define

$$\begin{aligned} f(x_1, x_2, \beta(\theta)) &= f^*(x_1, x_2, \beta(\theta)) - f^*(x_1, x_2, \beta(\theta_0)), \\ f_n(x_1, x_2, \beta(\theta)) &= f_n^*(x_1, x_2, \beta(\theta)) - f_n^*(x_1, x_2, \beta(\theta_0)), \end{aligned}$$

where

$$\begin{aligned} f^*(x_1, x_2, \beta(\theta)) &= \delta_2 I(v_1 \geq v_2) I(\beta^T z_1 - \beta^T z_2 > 0) + \delta_1 I(v_2 \geq v_1) I(\beta^T z_2 - \beta^T z_1 > 0), \\ f_n^*(x_1, x_2, \beta(\theta)) &= \delta_2 I(v_1 \geq v_2) s_n(\beta^T z_1 - \beta^T z_2) + \delta_1 I(v_2 \geq v_1) s_n(\beta^T z_2 - \beta^T z_1). \end{aligned}$$

Note that f_n is the smooth approximation of the function f . Write

$$\tau(x, \beta(\theta)) = E\{f^*(x, X, \beta(\theta))\}.$$

We now consider its smoothed counterpart $E\{\tau_n(x, \beta(\theta))\}$ with $\tau_n(x, \beta(\theta)) = E\{f_n^*(x, X, \beta(\theta))\}$ and the expectation over X . Denote P_n as the empirical measure and U_n as the U-process operator as in (2). Consider the function $\Gamma_n(\theta) = H_n(\theta) - H_n(\theta_0)$.

Denote $\nabla_1 \tau_n(x, \beta(\theta)) = \frac{\partial}{\partial \theta_1} \tau(x, \beta(\theta))$ and $\nabla_2 \tau_n(x, \beta(\theta)) = \frac{\partial^2 \tau_n(x, \beta(\theta))}{\partial \theta_1 \partial \theta_1^T}$. By arguments similar to those in (14), we have

$$\nabla_1 \tau_n(x, \beta(\theta)) = \nabla_1 \tau(x, \beta(\theta)) + O(\sigma_n) \text{ and } \nabla_2 \tau_n(x, \beta(\theta)) = \nabla_2 \tau(x, \beta(\theta)) + O(\sigma_n).$$

We can write $\Gamma_n(\theta) = \Gamma_{n0}(\theta) + P_n g_n(\cdot, \theta) + U_n h_n(\cdot, \theta)$, where

$$\begin{aligned} \Gamma_{n0}(\theta) &= E f_n(\cdot, \cdot, \theta), \\ g_n(x, \theta) &= E f_n(x, \cdot, \theta) + E f_n(\cdot, x, \theta) - 2\Gamma_{n0}(\theta) = \tau_n(x, \theta) - \tau_n(x, \theta_0) - 2\Gamma_{n0}(\theta), \text{ and} \\ h_n(x_1, x_2, \theta) &= f_n(x_1, x_2, \theta) - E f_n(x_1, \cdot, \theta) - E f_n(\cdot, x_2, \theta) + \Gamma_{n0}(\theta). \end{aligned}$$

It can be shown that

$$\begin{aligned} \Gamma_{n0}(\theta) &= \frac{1}{2}(\theta_1 - \theta_{10})^T A(\theta_1 - \theta_{10}) + O(\sigma_n |\theta_1 - \theta_{10}|) + o(|\theta_1 - \theta_{10}|^2), \\ P_n g_n(\cdot, \theta) &= n^{-1/2}(\theta_1 - \theta_{10})^T G_n + O(\sigma_n |\theta_1 - \theta_{10}|) + o(|\theta_1 - \theta_{10}|^2), \end{aligned}$$

where $A = -E\{\nabla_2 \tau(x, \beta(\theta_0))\}/2$, $G_n = \sqrt{n} P_n \partial \tau(\cdot, \theta_0) / \partial \theta_1 \rightarrow_d N(0, B)$, and $U_n h_n(\cdot, \cdot, \theta) = o_p(n^{-1})$ uniformly in an $o_p(1)$ neighborhood of θ_0 .

Thus, we have

$$\Gamma_n(\theta) = \frac{1}{2}(\theta - \theta_0)^T A(\theta - \theta_0) + n^{-1/2}(\theta_1 - \theta_{10})^T G_n + O(\sigma_n|\theta_1 - \theta_{10}|) + o(|\theta_1 - \theta_{10}|^2) + o_p(n^{-1}).$$

By a Taylor expansion,

$$\begin{aligned} & \sum_k \{p_\lambda(|\widehat{\theta}_{1k}|) - p_\lambda(|\theta_{10k}|)\} \\ &= \sum_k p'_\lambda(|\theta_{10k}|) \text{sgn}(\theta_{10k}) (\widehat{\theta}_{1k} - \theta_{10k}) + \sum_k \{p''_\lambda(|\theta_{10k}^*|) (\widehat{\theta}_{1k} - \theta_{10k})^2\} \\ &= D_1(\theta_{10})(\widehat{\theta}_1 - \theta_{10}) + (\widehat{\theta}_1 - \theta_{10})^T D_2(\theta_{10})(\widehat{\theta}_1 - \theta_{10}) + o\left\{\|\theta_1 - \theta_{10}\|^2\right\}, \end{aligned}$$

where

$$\begin{aligned} D_1(\theta_1) &= \text{diag}\{p'_\lambda(|\theta_{11}|) \text{sgn}(\theta_{11}), \dots, p'_\lambda(|\theta_{1d_1}|) \text{sgn}(\theta_{1d_1})\}, \text{ and} \\ D_2(\theta_{10}^*) &= \text{diag}\{p''_\lambda(|\theta_{11}|), \dots, p''_\lambda(|\theta_{1d_1}|)\}. \end{aligned}$$

Hence

$$\begin{aligned} Q_n(\theta) - Q_n(\theta_0) &= \frac{1}{2}(\theta_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\theta_1 - \theta_{10}) + (\theta_1 - \theta_{10})^T \{n^{-1/2}G_n - D_1(\theta_{10})\} \\ &\quad + O(\sigma_n \|\theta_1 - \theta_{10}\|) \\ &\quad + o(\|\theta_1 - \theta_{10}\|^2) \\ &\quad + o_p(n^{-1}), \end{aligned} \tag{B2}$$

Let $\tilde{\theta}_1 = \theta_{10} - \{A - 2D_2(\theta_{10})\}^{-1} \{n^{-1/2}G_n - D_1(\theta_{10})\}$ and $\tilde{\theta} = (\tilde{\theta}_1, 0)$. Note that $\tilde{\theta}_1$ maximizes the approximate quadratic form in (B2) and $\tilde{\theta}_1 - \theta_{10} = O_p(n^{-1/2})$ under condition (4). Apply (B2) at both $\tilde{\theta}$ and θ , we have

$$\begin{aligned} & Q_n(\widehat{\theta}) - Q_n(\theta_0) \\ &= \frac{1}{2}(\widehat{\theta}_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\widehat{\theta}_1 - \theta_{10}) + (\widehat{\theta}_1 - \theta_{10})^T \{n^{-1/2}G_n - D_1(\theta_{10})\} + o_p(n^{-1}) \\ &= \frac{1}{2}(\tilde{\theta}_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\tilde{\theta}_1 - \theta_{10}) - (\tilde{\theta}_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\tilde{\theta}_1 - \theta_{10}) + o_p(n^{-1}) \end{aligned}$$

and

$$\begin{aligned} & Q_n(\tilde{\theta}) - Q_n(\theta_0) \\ &= \frac{1}{2}(\tilde{\theta}_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\tilde{\theta}_1 - \theta_{10}) + (\tilde{\theta}_1 - \theta_{10})^T \{n^{-1/2}G_n - D_1(\theta_{10})\} + o_p(n^{-1}) \\ &= -\frac{1}{2}(\tilde{\theta}_1 - \theta_{10})^T \{A - 2D_2(\theta_{10})\}(\tilde{\theta}_1 - \theta_{10}) + o_p(n^{-1}). \end{aligned}$$

Subtract the first equation by the second one, we have

$$Q_n(\widehat{\theta}_{1n}) - Q_n(\tilde{\theta}_{1n}) = \frac{1}{2}(\widehat{\theta}_1 - \tilde{\theta}_1)^T \{A - 2D_2(\theta_{10})\}(\widehat{\theta}_{1n} - \tilde{\theta}_1) + o_p(n^{-1}).$$

Noting that $Q_n(\hat{\theta}) - Q_n(\tilde{\theta}_{1n}) \geq 0$ as $\hat{\theta}$ maximizes $Q_n(\theta)$ and $Q_n(\hat{\theta}) - Q_n(\tilde{\theta}_{1n}) \leq o_p(1)$ as $\tilde{\theta}_1$ maximizes the approximate quadratic form in (B2), together with that $\{A - 2nD_2(\theta_{10})\}$ is positive definite for large n , this implies that

$$0 \leq \left\{n^{-1/2}(\widehat{\theta}_1 - \tilde{\theta}_1)\right\}^T \{A - 2D_2(\theta_{10})\} \left\{n^{-1/2}(\widehat{\theta}_1 - \tilde{\theta}_1)\right\} \leq o_p(1).$$

Therefore

$$n^{-1/2}(\widehat{\theta}_1 - \tilde{\theta}_1) = o_p(1).$$

Since

$$-n^{1/2} \left[\{A - 2D_2(\theta_{10})\} \{\tilde{\theta}_1 - \theta_{10}\} + D_1(\theta_{10}) \right] = G_n \rightarrow N(0, B),$$

by Slutsky's theorem, we have

$$-n^{1/2} \left[\{A - 2D_2(\theta_{10})\} \{\widehat{\theta}_1 - \theta_{10}\} + D_1(\theta_{10}) \right] \rightarrow N(0, B).$$

Table 1

Simulation – variable selection: probability of correctly identifying zero versus nonzero regression coefficients.

	Cox	AFT	U-estimate	LASSO	aLASSO	bridge
Set I						
β_1	1.000	1.000	1.000	1.000	1.000	1.000
β_2	1.000	1.000	1.000	1.000	1.000	0.965
β_3	1.000	1.000	1.000	1.000	1.000	0.725
β_4	0.000	0.000	0.000	0.195	0.730	0.990
β_5	0.000	0.000	0.000	0.235	0.695	1.000
β_6	0.000	0.000	0.000	0.210	0.685	0.990
β_7	0.000	0.000	0.000	0.220	0.755	0.995
Set II						
β_1	1.000	1.000	1.000	1.000	1.000	1.000
β_2	1.000	1.000	1.000	1.000	1.000	0.995
β_3	1.000	1.000	1.000	1.000	1.000	0.870
β_4	0.000	0.000	0.000	0.380	0.930	1.000
β_5	0.000	0.000	0.000	0.385	0.945	1.000
β_6	0.000	0.000	0.000	0.390	0.930	1.000
β_7	0.000	0.000	0.000	0.370	0.925	1.000
Set III						
β_1	1.000	1.000	1.000	1.000	1.000	1.000
β_2	1.000	1.000	1.000	1.000	1.000	0.990
β_3	1.000	1.000	1.000	1.000	1.000	0.805
β_4	0.000	0.000	0.000	0.350	0.890	1.000
β_5	0.000	0.000	0.000	0.315	0.880	1.000
β_6	0.000	0.000	0.000	0.310	0.885	1.000
β_7	0.000	0.000	0.000	0.360	0.860	1.000

aLASSO, adaptive LASSO.

Table 2

Simulation – estimation: med (median) and mad (median absolute deviation).

		Cox		AFT		U-estimate		LASSO		aLASSO		bridge	
		med	mad	med	mad	med	mad	med	mad	med	mad	med	mad
Set I													
β_1	1	1.042	0.157	0.807	0.108	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_2	0.8	0.870	0.177	0.656	0.112	0.859	0.210	0.748	0.167	0.762	0.189	0.804	0.188
β_3	0.6	0.629	0.161	0.491	0.094	0.628	0.162	0.544	0.123	0.542	0.163	0.559	0.214
β_4	0	-0.010	0.140	-0.016	0.110	-0.035	0.148	0.000	0.089	0.000	0.000	0.000	0.000
β_5	0	-0.010	0.125	0.000	0.103	-0.005	0.159	0.000	0.089	0.000	0.000	0.000	0.000
β_6	0	-0.007	0.136	0.005	0.098	-0.013	0.184	0.000	0.092	0.000	0.000	0.000	0.000
β_7	0	-0.034	0.132	-0.033	0.106	-0.026	0.136	0.000	0.069	0.000	0.000	0.000	0.000
Set II													
β_1	1	13.013	1.772	0.998	0.016	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_2	0.8	10.375	1.322	0.801	0.014	0.827	0.022	0.804	0.020	0.814	0.026	0.814	0.029
β_3	0.6	7.802	1.035	0.599	0.015	0.617	0.024	0.601	0.023	0.605	0.027	0.601	0.031
β_4	0	0.026	0.187	-0.001	0.013	0.001	0.019	0.000	0.005	0.000	0.000	0.000	0.000
β_5	0	-0.015	0.199	0.000	0.014	-0.003	0.018	0.000	0.006	0.000	0.000	0.000	0.000
β_6	0	-0.006	0.209	-0.001	0.013	0.000	0.019	0.000	0.006	0.000	0.000	0.000	0.000
β_7	0	-0.003	0.209	0.000	0.014	0.001	0.020	0.000	0.005	0.000	0.000	0.000	0.000
Set III													
β_1	1	1.374	0.663	1.196	0.266	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
β_2	0.8	1.093	0.622	0.955	0.207	0.831	0.043	0.802	0.037	0.812	0.043	0.817	0.052
β_3	0.6	0.806	0.404	0.702	0.164	0.623	0.036	0.594	0.031	0.598	0.033	0.591	0.057
β_4	0	-0.006	0.258	-0.004	0.092	-0.001	0.030	0.000	0.010	0.000	0.000	0.000	0.000
β_5	0	0.040	0.265	0.006	0.100	0.001	0.032	0.000	0.014	0.000	0.000	0.000	0.000
β_6	0	-0.012	0.242	-0.002	0.092	-0.001	0.030	0.000	0.010	0.000	0.000	0.000	0.000
β_7	0	-0.042	0.245	0.005	0.100	0.001	0.027	0.000	0.010	0.000	0.000	0.000	0.000

aLASSO, adaptive LASSO.

Table 3

PBC data: estimation under different approaches.

	Cox	AFT	U-estimate	LASSO	aLASSO	bridge
Z_1	0.343 (1.000)	0.176 (1.000)	1.000	1.000	1.000	1.000
Z_2	-0.330 (-0.960)	-0.259 (-1.470)	-1.981	-0.849	-0.856	-2.583
Z_3	0.026 (0.076)	-0.002 (-0.011)	-0.522	-0.084	—	—
Z_4	-0.247 (-0.720)	-0.562 (-3.190)	-2.819	—	-2.610	—
Z_5	-0.417 (-1.214)	-0.236 (-1.337)	-3.112	-1.372	-1.677	-6.238
Z_6	-0.048 (0.140)	-0.033 (-0.188)	-0.277	-0.046	-0.018	—
Z_7	-0.725 (-2.113)	-0.425 (-2.412)	-1.543	-0.032	—	—
Z_8	-0.336 (-0.979)	-0.216 (-1.225)	-0.920	-0.016	—	—
Z_9	-0.093 (-0.270)	-0.005 (-0.026)	0.600	0.018	—	—
Z_{10}	-0.233 (-0.680)	-0.156 (-0.885)	-0.522	-0.414	-0.266	—
Z_{11}	0.341 (0.994)	0.274 (1.555)	1.526	0.188	0.620	—
Z_{12}	-0.210 (-0.612)	-0.155 (-0.880)	-0.574	—	-0.008	—
Z_{13}	-0.275 (-0.801)	-0.238 (-1.350)	-3.291	-0.684	-1.061	—
Z_{14}	0.034 (0.100)	0.029 (0.164)	-1.292	-0.271	-0.294	—
Z_{15}	0.098 (0.287)	0.072 (0.407)	-0.115	-0.008	—	—
Z_{16}	-0.224 (-0.652)	-0.184 (-1.044)	-1.749	-0.670	-0.701	—

aLASSO, adaptive LASSO. “—” indicates a zero estimate. To make estimates more comparable, for the Cox and AFT, the estimates are also re-scaled in the parentheses so that the first one is 1.

Table 4

UIS data: estimation under different approaches.

	Cox	AFT	U-estimate	LASSO	aLASSO	bridge
Z ₁	0.725 (1.000)	0.793 (1.000)	1.000	1.000	1.000	1.000
Z ₂	0.130 (0.179)	0.113 (0.143)	0.087	0.064	0.013	—
Z ₃	-0.050 (-0.068)	-0.058 (-0.073)	-0.065	-0.057	-0.030	-0.030
Z ₄	-0.144 (-0.199)	-0.142 (-0.179)	-0.069	-0.066	-0.043	—
Z ₅	0.306 (0.423)	0.362 (0.456)	0.201	0.203	0.206	0.265
Z ₆	-0.138 (-0.191)	-0.144 (-0.181)	-0.019	-0.006	—	—
Z ₇	-0.444 (-0.612)	-0.477 (-0.601)	-0.232	-0.165	-0.087	—
Z ₈	-0.002 (-0.003)	0.008 (0.010)	-0.032	—	—	—
Z ₉	-0.153 (-0.212)	-0.066 (-0.084)	-0.104	-0.038	-0.010	—
Z ₁₀	0.013 (0.017)	0.102 (0.129)	0.025	0.055	—	—
Z ₁₁	0.399 (0.550)	0.355 (0.448)	0.120	0.112	0.036	—
Z ₁₂	0.210 (0.290)	0.131 (0.165)	-0.037	—	—	—

aLASSO, adaptive LASSO. “—” indicates a zero estimate. To make estimates more comparable, for the Cox and AFT, the estimates are also re-scaled in the parentheses so that the first one is 1.