Published in final edited form as:

J Multivar Anal. 2018 November ; 168: 334–351. doi:10.1016/j.jmva.2018.08.007.

Broken adaptive ridge regression and its asymptotic properties

Linlin Dai^a, Kani Chen^b, Zhihua Sun^c, Zhenqiu Liu^d, and Gang Li^{e,*}

^aSouthwestern University of Finance and Economics, Chengdu, China

^bDepartment of Mathematics, Hong Kong University of Science and Technology, Hong Kong

^cInstitute of Mathematics, Ocean University of China, Qingdao, China

^dSamuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

^eDepartment of Biostatistics, School of Public Health, University of California at Los Angeles, CA 90095-1772, USA

Abstract

This paper studies the asymptotic properties of a sparse linear regression estimator, referred to as broken adaptive ridge (BAR) estimator, resulting from an L_0 -based iteratively reweighted L_2 penalization algorithm using the ridge estimator as its initial value. We show that the BAR estimator is consistent for variable selection and has an oracle property for parameter estimation. Moreover, we show that the BAR estimator possesses a grouping effect: highly correlated covariates are naturally grouped together, which is a desirable property not known for other oracle variable selection methods. Lastly, we combine BAR with a sparsity-restricted least squares estimator and give conditions under which the resulting two-stage sparse regression method is selection and estimation consistent in addition to having the grouping property in high- or ultrahigh-dimensional settings. Numerical studies are conducted to investigate and illustrate the operating characteristics of the BAR method in comparison with other methods.

Keywords

Feature selection; Grouping effect; L_0 -penalized regression; Oracle estimator; Sparsity recovery

1. Introduction

Simultaneous variable selection and parameter estimation is an essential task in statistics and its applications. A natural approach to variable selection is L_0 -penalized regression, which directly penalizes the cardinality of a model through well-known information criteria such as Mallows's C_p [35], Akaike's information criterion (AIC) [1], the Bayesian information criterion (BIC) [9, 38], and risk inflation criteria (RIC) [17]. It has also been shown to have an optimality property for variable selection and parameter estimation [40]. However, the L_0 -penalization problem is nonconvex and finding its global optima requires exhaustive

*Corresponding author vli@ucla.edu (Gang Li).

combinatorial best subset search, which is NP-hard and computationally infeasible even for data in moderate dimension. Moreover, it can be unstable for variable selection [6]. A popular alternative is L_1 -penalized regression, or Lasso [43], which is known to be consistent for variable selection [36, 48] although not consistent for parameter estimation [13, 51]. During the past two decades, much effort has been devoted to improving Lasso using various variants of the L_1 penalty, which are not only consistent for variable selection, but also consistent for parameter estimation [13, 15, 24, 27, 47, 51]. Yet another approach is to approximate the L_0 penalty using a surrogate L_0 penalty function, which could lead to a more sparse model with favorable numerical properties [11, 25, 40, 41].

This paper concerns a different L_0 -based approach to variable selection and parameter estimation which performs iteratively reweighted L_2 penalized regressions in order to approximate an L_0 penalized regression [19, 33]. Unlike a surrogate L_0 penalization method such as the truncated Lasso penalization (TLP) method [41], the L_0 -based iteratively reweighted L_2 penalization method can be viewed as performing a sequence of surrogate L_0 penalizations, where each reweighted L_2 penalty serves as an adaptive surrogate L_0 penalty and the approximation of L_0 penalization improves with each iteration. It is also computationally appealing since each iteration only involves a reweighted L_2 penalization. The idea of iteratively reweighted penalization has its roots in the well-known Lawson algorithm [29] in classical approximation theory, which later found applications in L_d minimization for values of $d \in (0, 1)$ [37], sparse signal reconstruction [21], compressive sensing [7, 8, 10, 20, 45], and variable selection [19, 33]. However, despite its increasing popularity in applications and favorable numerical performance, theoretical studies of the iteratively reweighted L_2 penalization method have so far focused on its non-asymptotic numerical convergence properties [10, 19, 33]. Its statistical properties such as selection consistency and estimation consistency remain unexplored.

The primary purpose of this paper is to study rigorously the asymptotic properties of a specific version of the L_0 -based iteratively reweighted L_2 -penalization method for the linear model. We will refer to the resulting estimator as the broken adaptive ridge (BAR) estimator to reflect the fact that it is obtained by fitting adaptively reweighted ridge regression which eventually breaks part of the ridge with the corresponding weights diverging to infinity to produce sparsity. The BAR algorithm uses an L_2 -penalized (or ridge) estimator as its initial estimator and updates the estimator iteratively using reweighted L_2 penalizations where the weight for each coefficient at each iteration step is the inverse of its squared estimate from the previous iteration as defined later in Eqs. (2)–(3) of Section 2.

Our key theoretical contribution is to establish two large-sample properties for the BAR estimator. First, we will show that the BAR estimator possesses an oracle property: when the true model is sparse, with probability tending to 1, it estimates the zero coefficients as zeros and estimates the non-zero coefficients as well as the scenario when the true sub-model is known in advance. To this end, we stress that because the BAR estimator is the limit of an iterative penalization algorithm, commonly used techniques for deriving oracle properties of a single-step penalized regression method do not apply directly. As pointed out by a reviewer, the common strategies used in analyzing approximate solutions to nonconvex penalized method such as SCAD and MCP usually rely on an initial estimator that is close to

the optimal, and then show that some kind of iterative refinement of the initial estimator would lead to a good local estimator [15, 34]. Our derivations of the asymptotic oracle properties for BAR are inspired by this idea, but the technical details are substantially different from those for other variable selection methods in the literature. A unique feature of the BAR algorithm is that at each reweighted L_2 iteration, the resulting estimator is not sparse. We show that with a good initial ridge estimator, each BAR iteration shrinks the estimates of the zero coefficients towards zero and drives the estimates of the nonzero coefficients towards an oracle ridge estimator. Consequently, sparsity and selection consistency are achieved only in the limit of the BAR algorithm and the estimation consistency of the ridge estimator for the nonzero coefficients is preserved. Second, we will show that the BAR estimator has a grouping effect that highly correlated variables are naturally grouped together with similar coefficients, which is a desirable property inherited from the ridge estimator. Therefore, our results show that the BAR estimator enjoys the best of the L_0 and L_2 penalized regressions, while avoiding some of their limitations. Lastly, we consider high- or ultrahigh-dimensional settings when the dimension is larger or much larger than the sample size. Frommlet and Nuel [19] suggested to use the mBIC penalty for the BAR method in high-dimensional settings, but its asymptotic consistency has yet to be studied rigorously. We propose a different extension of the BAR method to ultrahigh-dimensional settings by combining it with a sparsity-restricted least squares estimation method and show rigorously that the resulting two-stage estimator is asymptotically consistent for variable selection and parameter estimation, and that it retains the grouping property.

In Section 2, we review the BAR estimator. In Section 3, we state our main results on its oracle property and grouping property for the general non-orthogonal case; we also develop rigorously a two-stage BAR method for high- or ultrahigh-dimensional settings. Section 4 illustrates the finite-sample operating characteristics of BAR in comparison with some popular methods for variable selection, parameter estimation, prediction, and grouping. Section 5 presents illustrations involving a prostate cancer data and a gut microbiota data. Concluding remarks are given in Section 6. Proofs of the theoretical results are deferred to the Appendix.

2. Broken adaptive ridge estimator

Consider the linear regression model

$$\mathbf{y} = \sum_{j=1}^{p_n} \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon},$$

where $\mathbf{y} \in \mathbb{R}^n$ is a response vector, $\mathbf{x}_1, \dots, \mathbf{x}_{p_n} \in \mathbb{R}^n$ are predictors, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p_n})^\top$ is the vector of regression coefficients. Assume that $\boldsymbol{\beta}_0$ is the true parameter. Without loss of generality, we suppose $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{p_n}) = (\mathbf{w}_1, \dots, \mathbf{w}_n)^\top$ and $\mathbf{y}$ are centralized. In this paper, $\|\cdot\|$ represents the Euclidean norm for a vector and spectral norm for a matrix. Let

$$\hat{\beta}^{(0)} = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \xi_n \sum_{j=1}^{p_n} \beta_j^2 \right\} = (X^T X + \xi_n I)^{-1} X^T y \quad (1)$$

be the ridge estimator, where $\xi_n > 0$ is a tuning parameter. For any integer $k \geq 1$, define

$$\hat{\beta}^{(k)} = g\{\hat{\beta}^{(k-1)}\}, \quad (2)$$

where

$$g(\tilde{\beta}) = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda_n \sum_{j=1}^{p_n} \beta_j^2 / \tilde{\beta}_j^2 \right\} = \{X^T X + \lambda_n D(\tilde{\beta})\}^{-1} X^T y, \quad (3)$$

and $D(\tilde{\beta}) = \text{diag}(\tilde{\beta}_1^{-2}, \dots, \tilde{\beta}_{p_n}^{-2})$. The broken adaptive ridge (BAR) estimator is defined as

$$\hat{\beta}^* = \lim_{k \rightarrow \infty} \hat{\beta}^{(k)}.$$

Remark 1. [Weight selection and arithmetic overflow] The initial ridge estimator $\hat{\beta}^{(0)}$ and its subsequent updates $\hat{\beta}^{(k)}$ do not yield any zero coefficient. Thus, the weights $\tilde{\beta}_i^{-2}$ s in each iteration are well defined. However, as the iterated values converge to its limit, the weight matrix $D\{\hat{\beta}^{(k-1)}\}$ will inevitably run into division by a very small nonzero value, which could cause an arithmetic overflow. A commonly used solution to this problem is to introduce a small perturbation, or more specifically, replace $D(\tilde{\beta}) = \text{diag}(\tilde{\beta}_1^{-2}, \dots, \tilde{\beta}_{p_n}^{-2})$ by $\text{diag}\{1/(\tilde{\beta}_1^2 + \delta^2), \dots, 1/(\tilde{\beta}_{p_n}^2 + \delta^2)\}$ for some $\delta > 0$ to avoid numerical instability [10, 19].

Here we provide a different solution by rewriting formula (3) as

$$g(\tilde{\beta}) = \Gamma(\tilde{\beta}) \left\{ \tilde{X}^T(\tilde{\beta}) \tilde{X}(\tilde{\beta}) + \lambda_n I_{p_n} \right\}^{-1} \tilde{X}^T(\tilde{\beta}) y, \quad (4)$$

where $\Gamma(\tilde{\beta}) = \text{diag}(\tilde{\beta}_1, \dots, \tilde{\beta}_{p_n})$, and $\tilde{X}(\tilde{\beta}) = X\Gamma(\tilde{\beta})$. Because the right-hand side of (4) involves only multiplication, instead of division, by the $\tilde{\beta}_j$ s, numerical stability is achieved without having to introduce a perturbation δ . We also note that the BAR algorithm based on (4) is equivalent to the algorithm of Liu and Li [33].

Remark 2. [Orthogonal case] To appreciate how the BAR method works, it is helpful to examine the orthogonal case which admits a closed form for the BAR estimator and which has been previously studied in [19]. Assume that $\mathbf{X}^\top \mathbf{X}/n = \mathbf{I}_{p_n}$. Then, for each $j \in \{1, \dots, p_n\}$, the j th component of $g(\tilde{\beta})$ defined by (3) is equal to

$$g_j(\tilde{\beta}_j) = \frac{\tilde{\beta}_j^2}{\tilde{\beta}_j^2 + \lambda_n/n} \tilde{\beta}_j^{(\text{OLS})},$$

where $\hat{\beta}^{(\text{OLS})} = n^{-1} \mathbf{X}^\top \mathbf{y}$ is the ordinary least squares (OLS) estimate. Hence the BAR algorithm can be implemented component by component. First, we show the existence and uniqueness of the BAR estimator $\hat{\beta}^*$. Without loss of generality, assume $\hat{\beta}_j^{(\text{OLS})} > 0$ and then $0 \leq \hat{\beta}_j^{(k)} \leq \hat{\beta}_j^{(\text{OLS})}$ for every integer k . Also note that the map $x \mapsto g(x) = x^2 \hat{\beta}_j^{(\text{OLS})} / (x^2 + \lambda_n/n)$ is increasing in x on $(0, \infty)$. Set $\phi(x) = g(x) - x$. For fixed n and λ_n the following facts can be verified easily.

- (i) If $\hat{\beta}_j^{(\text{OLS})} < 4\lambda_n/n$, then $\phi(x) < 0$. Therefore, $\hat{\beta}_j^{(k+1)} < \hat{\beta}_j^{(k)}$ for all k and $\hat{\beta}_j^* = 0$.
- (ii) If $\hat{\beta}_j^{(\text{OLS})} = 4\lambda_n/n$, then $\phi(x) = 0$. This leads to $\hat{\beta}_j^{(k+1)} \leq \hat{\beta}_j^{(k)}$ for all k . Then,

$$\hat{\beta}_j^* = \begin{cases} 0 & \text{if } \hat{\beta}_j^{(0)} < \hat{\beta}_j^{(\text{OLS})}/2, \\ \hat{\beta}_j^{(\text{OLS})}/2 & \text{otherwise.} \end{cases}$$

- (iii) If $\hat{\beta}_j^{(\text{OLS})} \geq 4\lambda_n/n$, then $\phi(x) = 0$ when $x \in [x_1, x_2]$, where $x_1 = \hat{\beta}_j^{(\text{OLS})}/2 - \left\{ \hat{\beta}_j^{(\text{OLS})}/4 - \lambda_n/n \right\}^{1/2}$ and $x_2 = \hat{\beta}_j^{(\text{OLS})}/2 + \left\{ \hat{\beta}_j^{(\text{OLS})}/4 - \lambda_n/n \right\}^{1/2}$ and $\phi(x) < 0$ otherwise. As a result,

$$\hat{\beta}_j^* = \begin{cases} 0 & \text{if } \hat{\beta}_j^{(0)} < x_1, \\ \hat{\beta}_j^{(\text{OLS})}/2 + \sqrt{\hat{\beta}_j^{(\text{OLS})}/4 - \lambda_n/n}, & \text{otherwise.} \end{cases}$$

In conclusion, the BAR estimator always converges to a unique solution depending on the settings of the initial value. Since we choose the ridge estimator as the initial value, which would be close to the OLS estimator under mild conditions, the BAR algorithm for the j th component converges to the following limit:

$$\hat{\beta}_j^* = \begin{cases} 0 & \text{if } \hat{\beta}_j(\text{OLS}) \in (-2\sqrt{\lambda_n/n}, 2\sqrt{\lambda_n/n}), \\ \hat{\beta}_j(\text{OLS})/2 + \sqrt{\hat{\beta}_j^2(\text{OLS})/4 - \lambda_n/n} & \text{if } \hat{\beta}_j(\text{OLS}) \notin (-2\sqrt{\lambda_n/n}, 2\sqrt{\lambda_n/n}). \end{cases} \quad (5)$$

Hence, the BAR estimator $\hat{\beta}^*$ shrinks the small values of $\hat{\beta}(\text{OLS})$ to 0 and keeps its large values essentially unchanged provided that λ_n/n is small. Figure 1 depicts the BAR thresholding function (5), along with some well-known thresholding functions (L_0 , Lasso, Adaptive Lasso, and SCAD), suggesting that BAR gives a good approximation to an L_0 penalized regression. The existence and uniqueness of the BAR estimator for $\mathbf{X}$ in a general form are nontrivial, although they have been verified empirically in our extensive numerical studies. A slightly weaker theoretical result is provided in Theorem 1(i) of Section 3. Therefore, BAR is expected to behave like an L_0 -penalized regression, but is more stable and easily scalable to high-dimensional settings.

3. Large-sample properties of BAR

3.1. Oracle property

We next develop the oracle properties of the BAR estimator for the general case without requiring the design matrix to be orthogonal.

Let $\beta_0 = (\beta_{01}, \dots, \beta_{0p_n})^\top$ be the true parameter value with dimension p_n diverging to infinity

but $p_n < n$. For simplicity, write $\beta_0 = (\beta_{01}^\top, \beta_{02}^\top)^\top$ where $\beta_{01} \in \mathbb{R}_n^q$ and $\beta_{02} \in \mathbb{R}^{p_n - q_n}$.

Assume $\beta_{01} \neq 0$ and $\beta_{02} = 0$. Let $\hat{\beta}^* = (\hat{\beta}_{01}^{*\top}, \hat{\beta}_{02}^{*\top})^\top$ denote the BAR estimator of β_0 . Set

$\mathbf{X}_1 = (\mathbf{x}_1, \dots, \mathbf{x}_{q_n})$, $\Sigma_{n1} = \mathbf{X}_1^\top \mathbf{X}_1/n$ and $\Sigma_n = \mathbf{X}^\top \mathbf{X}/n$.

The following conditions are needed to derive the asymptotic properties of the BAR estimator.

- (C1) The errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed random variables with mean zero and variance $0 < \sigma^2 < \infty$.
- (C2) There exists a constant $C > 1$ such that $0 < 1/C < \lambda_{\min}(\Sigma_n) \leq \lambda_{\max}(\Sigma_n) < C < \infty$ for every integer n .
- (C3) Let $a_{0n} = \min_{1 \leq j \leq q_n} |\beta_{0j}|$ and $a_{1n} = \min_{1 \leq j \leq q_n} |\beta_{0j}|$. As $n \rightarrow \infty$, $p_n q_n / \sqrt{n} \rightarrow 0$, $\xi_n a_{1n} / \sqrt{n} \rightarrow 0$, $(p_n/n)^{1/2} / a_{0n} \rightarrow 0$, and $\lambda_n a_{1n} (q_n/n)^{1/2} / a_{0n}^2 \rightarrow 0$.

Theorem 1 (Oracle property). *Assume the conditions (C1)–(C3) are satisfied. For any q_n -vector $\mathbf{b}_n$ satisfying $\|\mathbf{b}_n\| \leq 1$, define $s_n^2 = \sigma^2 \mathbf{b}_n^\top \Sigma_{n1}^{-1} \mathbf{b}_n$. Then, the fixed point of*

$f(\alpha) = \{\mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\alpha)\}^{-1} \mathbf{X}_1^\top \mathbf{y}$ exists and is unique, where $\mathbf{D}_1(\alpha) = \text{diag}(\alpha_1^{-2}, \dots, \alpha_{p_n}^{-2})$

Moreover, with probability tending to 1,

- (i) the BAR estimator $\hat{\beta}^* = (\hat{\beta}_1^{*\top}, \hat{\beta}_2^{*\top})^\top$ exists and is unique, where $\hat{\beta}_2^* = 0$ and $\hat{\beta}_1^*$ is the unique fixed point of $f(\alpha)$;
- (ii) $\sqrt{n} s_n^{-1} \mathbf{b}_1^\top (\hat{\beta}_1^* - \beta_{01}) \rightsquigarrow \mathcal{N}(0, 1)$.

3.2. Grouping effect

When the true model has a natural group structure, it is desirable to have all coefficients within a group clustered together. For example, by virtue of gene network relationships, some of the genes will strongly correlate and are referred to as grouped genes [39]. The next theorem shows that the BAR method has a grouping property.

Theorem 2. Assume that the columns of matrix $\mathbf{X}$ are standardized such that $x_{1j} + \dots + x_{nj} = 0$ and $\|\mathbf{x}_j\|_2^2 = 1$ for all $j \in \{1, \dots, p_n\}$. Let $\hat{\beta}^*$ be the BAR estimator. Then, with probability tending to 1, for any $i < j$,

$$|\hat{\beta}_i^* - 1 - \hat{\beta}_j^* - 1| \leq \frac{1}{\lambda_n} \|\mathbf{y}\| \sqrt{2(1 - \rho_{ij})},$$

provided that $\hat{\beta}_i^* \hat{\beta}_j^* \neq 0$, where $\rho_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ is the sample correlation of $\mathbf{x}_i$ and $\mathbf{x}_j$.

The above grouping property of BAR is similar to that of Elastic Net. In particular, Zou and Hastie [52] showed that if $\hat{\beta}_i(\text{EN}) \hat{\beta}_j(\text{EN}) > 0$, then

$$|\hat{\beta}_i(\text{EN}) - \hat{\beta}_j(\text{EN})| \leq \frac{1}{\lambda} \|\mathbf{y}\| \sqrt{2(1 - \rho_{ij})},$$

where $\hat{\beta}_i(\text{EN})$ is the Elastic Net estimate of β_i and λ is a tuning parameter. As seen in the proof of Theorem 2, the reason for stating the grouping effect of the BAR estimator in terms of $\hat{\beta}_i^{-1}$ is the weight $\hat{\beta}_i^{-2}$ used in the reweighted L_2 penalization.

3.3. High-dimensional setting

Like most oracle variable selection methods, the oracle property of the BAR estimator in Theorem 1 is derived for $p_n < n$. In high- or ultrahigh-dimensional problems where the number of variables p_n can be much larger than the sample size n , it is not clear if BAR would still enjoy the oracle property under the similar conditions. A commonly used strategy for ultrahigh-dimensional problems is to add a screening step to reduce the dimension before performing variable selection and parameter estimation [14]. During the past decades, many variable screening procedures have been developed with the sure screening property that with probability tending to 1, the screened model retains all important variables; see, e.g.,

[12, 14, 15, 22, 28, 30–32, 44, 46, 49, 50] and the references therein. Our results in Theorems 3 and 4 guarantee that when combined with a sure screening procedure, the BAR estimator has the oracle and grouping property in ultrahigh-dimensional settings.

Inspired by Xu and Chen [46], who studied a sparse maximum likelihood estimation method for generalized linear models, we introduce below a sparsity-restricted least squares estimate for screening joint effects for the linear model with unspecified error distribution and establish its sure screening property. Specifically, for a given k , define the following sparsity-restricted least squares estimate (sLSE):

$$\tilde{\beta}_{\text{sLSE}} = \arg \min_{\beta} \{ \ell(\beta) : \|\beta\|_0 \leq k \}, \quad (6)$$

where $\ell(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$. Solving for $\hat{\beta}_{\text{sLSE}}$ can be computationally demanding even for moderately large k since it is equivalent to the best subset search. Here we adopt the iterative hard-thresholding (IHT) algorithm of Blumensath and Davies [4] to approximately solve the optimization problem (6). Define a local quadratic approximation of $\ell(\beta)$ by

$$h(\gamma, \beta) = \ell(\beta) + (\gamma - \beta)^\top S(\beta) + \frac{u}{2} \|\gamma - \beta\|_2^2,$$

where $S(\beta) = \ell'(\beta) = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta)$ and $u > 0$ is a scaling parameter. The IHT algorithm solves (6) via the following iterative algorithm

$$\beta^{(t+1)} = \arg \min_{\gamma} \{ h(\gamma, \beta^{(t)}) : \|\gamma\|_0 \leq k \} \quad (7)$$

where $\beta^{(0)} = 0$. Note that without the sparsity constraint, (7) has an explicit solution

$$\gamma = \beta^{(t)} + u^{-1} 2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta^{(t)}).$$

Adding the sparsity constraint, the solution of (7) is then given by

$$\beta^{(t+1)} = \left(\gamma_1 \mathbf{1}_{\{|\gamma_1| \geq r\}}, \dots, \gamma_{p_n} \mathbf{1}_{\{|\gamma_{p_n}| \geq r\}} \right)^\top,$$

with r being the k th largest component of $|\gamma| = (|\gamma_1|, \dots, |\gamma_{p_n}|)^\top$. Therefore the update at each iteration of the IHT algorithm is straightforward and The following theorem shows that with an appropriate scaling parameter u , the sequence $\{\beta^{(t)}\}$ obtained from the above IHT algorithm decreases the objective function ℓ at every iteration and converges to a local solution of (6).

Theorem 3. Let $\{\beta^{(t)}\}$ be the sequence defined by (7). Assume that $u \geq 2\lambda_{\max}(\mathbf{X}^T \mathbf{X})$ where $\lambda_{\max}(\mathbf{X}^T \mathbf{X})$ denotes the largest eigenvalue of $\mathbf{X}^T \mathbf{X}$. Then

- (a) $\|\beta^{(t)}\|_0 \leq \|\beta^{(t+1)}\|_0$ for every integer t .
- (b) Assume that the parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$ and contains the true parameter value β_0 . If $\sum_S \mathbf{w}_S \mathbf{w}_S^T > 0$ for all submodels s such that $\|s\|_0 \leq k$, then $\{\beta^{(t)}\}$ converges to a local solution to (6).

Let $s_0 = \{j : \beta_{0j} \neq 0\}$ denote the true model and $\hat{s}_{\text{sLSE}} = \{j : \hat{\beta}_{\text{sLSE},j} \neq 0\}$ the screened model via the sparsity-restricted least squares estimate $\hat{\beta}_{\text{sLSE}}$. The following conditions are needed to establish the sure screening property of $\hat{s}_{\text{sLSE}}$.

- (D1) There exist $\omega_1, \omega_2 > 0$ and some nonnegative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$, $\min_{j \in s_0} |\beta_{0j}| \geq \omega_1 n^{-\tau_1}$ and $q_n < k - \omega_2 n^{\tau_2}$.
- (D2) $\ln(p_n) = O(n^d)$ for some $d \in [0, 1 - 2\tau_1 - 2\tau_2]$.
- (D3) There exist constants $c_1 > 0, \delta_1 > 0$ such that for sufficiently large n , $\lambda_{\min}(n^{-1} \mathbf{X}_s^T \mathbf{X}_s) \geq c_1$ for $\beta_s \in \{\beta_s : \|\beta_s - \beta_{s_0}\|_2 \leq \delta_1\}$ and $s \in S_+^{2k} = \{s : s_0 \subset s, \|s\|_0 \leq 2k\}$ where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix.
- (D4) Let $\sigma_i^2 = \text{var}(y_i)$. There exist positive constants c_2, c_3, c_4, σ such that $|x_{ij}| \leq c_2, |\mathbf{w}_i^T \beta_0| \leq c_3, |\sigma_i| \leq \sigma$ and for sufficiently large n ,

$$\max_{1 \leq j \leq p} \max_{1 \leq i \leq n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 \sigma_i^2} \right\} \leq c_4 n^{-1}.$$

Theorem 4. (Sure Joint Screening Property) Under conditions (D1)–(D4), $\Pr(s_0 \subset \hat{s}_{\text{sLSE}}) \rightarrow 1$ as $n \rightarrow \infty$.

Finally, let $\hat{\beta}_{\text{sLSE} - \text{BAR}}$ be BAR estimator for the screened submodel

$$\mathbf{y} = \sum_{j \in \hat{s}_{\text{sLSE}}} \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon}. \quad (8)$$

It then follows immediately from Theorems 1, 2 and 4 that under conditions (D1)–(D4), $\hat{\beta}_{\text{sLSE} - \text{BAR}}$ has the oracle properties for variable selection and parameter estimation and the grouping property for highly correlated covariates, provided that Model (8) satisfies conditions (C1)–(C3).

4. Simulation study

4.1. Variable selection, estimation, and prediction

We first present a simulation to illustrate the performance of BAR for variable selection, parameter estimation, and prediction in comparison with some L_1 -based variable selection methods (Lasso, adaptive Lasso, SCAD and MCP) and a surrogate L_0 method (TLP) [40] under some low-dimensional settings when $p = n$. Here we consider two linear models, viz.

$$\text{Model 1: True parameter } \beta_0 = (2, -3, 0, 0, 4, 0, \dots, 0)^T,$$

$$\text{Model 2: True parameter } \beta_0 = (2, -3, 0, 0, 4, 0.2, -0.3, 0, 0, 0.4, 0, \dots, 0)^T,$$

where in both models, the p -dimensional covariate vector $\mathbf{X}$ is generated from a multivariate Gaussian distribution with mean 0 and variance-covariance matrix $\Sigma = (r^{i-j})$, and the error $\mathbf{e}$ is Gaussian and independent of the covariates. Model 1 has sparse strong signals and Model 2 contains a mix of both strong and weak signals. Five-fold CV was used to select tuning parameters for each method. For BAR, we used ten grid points evenly spaced on the interval $[10^{-4}, 100]$ for ξ_n and ten equally spaced grid points on $[1, 10]$ for λ_n . We observed from our limited experience that BAR is insensitive to the choice of ξ_n and its solution path is essentially flat over a broad range of ξ_n . So the selection of λ_n is more relevant. In this simulation, we used only ten grid points to reduce the computational burden. In practice, however, a more refined grid for λ_n would generally be preferred. We used the R package `glmnet` [18] for Lasso, adaptive Lasso and `ncvreg` [5] for SCAD and MCP in our simulations. The TLP method [40] requires to specify a resolution tuning parameter τ . To illustrate the impact of τ , we considered two τ values (0.15 and 0.5), although a data-driven τ would be used in practice as demonstrated in [40]. We considered a variety of scenarios by varying n , p , and r for each of the above two models, but only present partial results in Table 1 due to space limitation.

From Table 1, it appears that the BAR and TLP methods generally produce more sparse models than the L_1 -based methods, with BAR having the lowest number of false positives (nonzeros). When the signals are strong (Model 1), BAR demonstrates better performance in most categories with (i) better variable selection performance (the lowest number of misclassifications (non-zeros and zeros), lowest false positives, highest probability of selecting the true model (TM), and equally low false rate (close to 0)); (ii) better estimation performance (the smallest estimation bias, MAB); and (iii) for equally good prediction performance (similar mean squared prediction error (MSPE)).

Similar trends are observed for Model 2, in which there are both strong and weak signals. BAR still produces the sparsest model with the lowest positive false positive rate and lowest number of misclassifications, but it tends to miss the small effects with more false negatives. For estimation, BAR has comparable or smaller bias, but no method dominates others across all scenarios. For prediction, all methods had comparable performance with similar MSPE.

Because some signals are weak and the sample size $n = 100$ is relatively small, it is challenging for all methods to select the exact correct model (TM close to zero), but for different reasons: BAR tends to miss some small signals with more false negatives (zeros) whereas the other methods tend to include some zeros as signals with more false positives. Moreover, as a surrogate L_0 method, the TLP method generally produces a more sparse model than an L_1 -based method. In principle, the TLP method with a smaller τ value gives a closer approximation to L_0 penalization and thus leads to a more sparse model as shown in Table A.1. In practice, however, τ cannot be too small since the TLP estimate obtained from Algorithm 1 in [40] would be no more sparse than its initial Lasso estimator.

In summary, BAR shows an advantage of more accurately selecting and estimating strong signals. On the other hand, if weak signals are also of interest and the sample size is relatively small, then an L_1 -based method with smaller false negatives might be preferred.

4.2. Grouping effect

We now present a simulation to illustrate the grouping effect of BAR for highly correlated covariates. Assume $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ with $p_n = 9$ and $\boldsymbol{\beta}_0 = (2, 2, 2, 0.4, 0.4, 0.4, 0, 0, 0)$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, 0.2^2 \mathbf{I}_n)$ and

$$\begin{cases} \mathbf{x}_i = \mathbf{z}_1 + \mathbf{e}_i, \mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, 30^2 \mathbf{I}_n) & \text{for } i \in \{1, 2, 3\}, \\ \mathbf{x}_i = \mathbf{z}_2 + \mathbf{e}_i, \mathbf{z}_2 \sim \mathcal{N}(\mathbf{0}, 30^2 \mathbf{I}_n) & \text{for } i \in \{4, 5, 6\}, \\ \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, 30^2 \mathbf{I}_n) & \text{for } i \in \{7, 8, 9\}, \end{cases}$$

where $\mathbf{e}_i$ are independent $\mathcal{N}(\mathbf{0}, 0.1^2 \mathbf{I}_n)$. In this setting, $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ form a group with latent factor $\mathbf{z}_1$, and $\mathbf{x}_5, \mathbf{x}_6$, and $\mathbf{x}_7$ form another group with latent factor $\mathbf{z}_2$. The within-group correlations are almost 1 and the between-group correlations are almost 0. Figure 2 plots the solution paths of BAR, Elastic Net, Lasso, Adaptive Lasso, SCAD, and MCP for a random sample of size $n = 200$. The estimated coefficients and cross-validation errors together with the corresponding optimal tuning parameters are provided in Table 2.

It is seen from Figure 2 and Table 2 that both BAR and Elastic Net selected all three variables in each of the two non-zero groups with almost evenly distributed coefficients within each group, whereas Lasso, adaptive Lasso, SCAD, and MCP failed to reveal any grouping information with disorganized paths and erratic coefficient estimates.

4.3. High-dimensional settings ($p \gg n$)

We finally present a simulation to illustrate the performance of BAR in comparison with some other variable selection methods when combined with the sparsity-restricted least squares estimate-based screening method in a high-dimensional setting. Specifically, we consider Models 1 and 2 in Section 3.1, but increase p to 2000. The simulation results for the two-stage estimators with $n = 100$ are presented in Table 3.

It is seen that for both Models 1 and 2, sLSE-BAR outperformed other methods for variable selection and parameter estimation with a lower number of misclassifications, fewer false positives, and a smaller estimation bias (MAB). Consistent with the low-dimensional settings, BAR shows similar prediction performance to the other methods. Again it is challenging for all methods to select exactly the true model under Model 2 when there are both strong and weak signals, but our limited simulations (not presented here) showed that TM (the probability of selecting the correct model) will improve, especially for BAR, as n increases.

5. Data examples

5.1. Prostate cancer data

As an illustration, we analyze the prostate cancer data described in [42]. There are eight clinical predictors, namely $\ln(\text{cancer volume})$ (lcavol), $\ln(\text{prostate weight})$ (lweight), age, $\ln(\text{amount of benign prostatic hyperplasia})$ (lpbh), seminal vesicle invasion (svi), $\ln(\text{capsular penetration})$ (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45). We consider the problem of selecting important predictors for the logarithm of prostate-specific antigen (lpsa). We randomly split the data set into a training set of size 67 and a test set of size 30, and compare the results of six variable selection methods: BAR, lasso, adaptive lasso, SCAD, MCP and Elastic Net by selecting nonzero features based on the training data set and test error as the mean squared prediction error on the test data set. We performed an analysis with $p = 36$ features consisting of all the main effects and two-way interactions. The five-fold CV method was used to select the optimal ξ_n and λ_n values for BAR, with 100 evenly-spaced grid points on the interval $[10^{-4}, 100]$ for ξ_n and 100 equally-spaced grid points on $[1, 100]$ for λ_n . The estimated coefficients are given in Table 4, together with a summary of the number of selected features and test errors for each method.

It is seen that BAR produces the sparsest model with three nonzero features as compared with at least six nonzero features by the other methods. Its prediction performance is comparable to other methods with a slightly lower test error. This is consistent with the simulation results in Section 4.1.

5.2. Gut microbiota data

In this example, we illustrate our method on a metagenomic dataset on gut microbiota that was downloaded from dbGaP under study ID phs000258 [53]. After aligning the 16S rRNA sequences of gut microbiota to reference sequences and taxonomy databases, we had a total of 240 taxa at the genus level. Here we are interested in identifying BMI associated taxa. Because taxa with small reads are not reliable and subject to noise and measurement error, we first dropped the taxa with average reads below 2, resulting in 62 features. For each taxon, we created a dummy variable $\mathbf{1}(\text{value of taxon} > 25\text{th quantile of the taxon})$. We used 304 patients after removing missing values. As in the previous example, we randomly split the data set into a training set of size 228 to select nonzero features and a test set of size 76 to compute the mean squared prediction error (MSPE). Table 5 gives the selected taxa, together with a summary of the number of selected features and test errors for each of six methods: BAR, lasso, adaptive lasso, SCAD, MCP and Elastic Net.

Consistent with the simulation results in Section 3, BAR yields the sparsest model (four nonzero features as compared to at least six nonzero features by the other methods) with similar test errors to other methods. Among the four identified taxa by BAR, *Faecalibacterium* is well known to be associated with obesity. For example, some recent studies showed that *Faecalibacterium* increases significantly in lean people and has a negative association with BMI [2, 23]. The other three taxa (*Collinsella*, *Megasphaera*, and *Turicibacter*) have also been shown to be correlated with obesity and BMI in animal models and clinical trials [3, 16, 26].

6. Discussion

We have established an oracle property and a grouping property for the BAR estimator, a sparse regression estimator resulted from an L_0 -based iteratively reweighted L_2 -penalized regressions using a ridge estimator as the initial value. Hence, BAR enjoys the best of both L_0 and L_2 penalized regressions while avoiding their pitfalls. For example, BAR shares the consistency properties of L_0 -penalized regression for variable selection and parameter estimation, avoids its instability, and is computationally more scalable to high-dimensional settings. BAR can also be viewed as a sparse version of the ridge estimator obtained by shrinking the small ridge coefficients to zeros via an approximate L_0 -penalized regression and consequently inherits the stability and grouping effect of the ridge estimator. In addition to the low-dimensional setting ($p < n$), we have extended the BAR method to high- or ultrahigh-dimensional settings when p is larger and potentially much larger than n , by combining it with a sparsity-restricted least squares estimate and provided conditions under which the two-stage estimator possesses the oracle property and grouping property.

We have conducted extensive simulations to study the operating characteristics of the L_0 -based BAR method in comparison with some popular L_1 -based variable selection methods. It was observed that as an L_0 -based method, BAR tends to produce a more sparse model with lower false positive rate and comparable prediction performance, and generally can select and estimate strong signals more accurately with a lower misclassification rate and a smaller estimation bias. In contrast, it is more likely to shrink weak signals to null when the sample size is not very large. Therefore, in practice, one may choose between an L_1 -based method and an L_0 -based method such as BAR based on whether or not weak signals are of interest.

Our theoretical developments for the asymptotic properties of the L_0 -based BAR estimator can be easily modified to analyze large sample properties of an L_d -based BAR estimator obtained by replacing $\tilde{\beta}_j^{-2}$ with $|\tilde{\beta}_j|^{d-2}$ in Eq. (3) for any $d \in [0, 1]$. Our simulations (not reported in this version of the paper) suggests that there is a trade-off between the false positive rate and false negative rate as d varies. Specifically, as d increases from 0 to 1, the L_d -based BAR estimator becomes less sparse with an increased false positive rate but decreased false negative rate.

Finally, this paper focuses on the linear model. Extensions to more complex models and applications such as generalized linear models and survival data are currently under investigation by our team.

Acknowledgments

The author thanks the Editor-in-Chief, an Associate Editor and two anonymous referees for their constructive and insightful comments and suggestions that have led to significant improvements of the paper. The research of Gang Li was partly supported by National Institute of Health Grants P30 CA-16042, UL1TR000124-02, and P50 CA211015.

Appendix A. Proofs of the theorems

To prove Theorems 1–4, we need to introduce some notations and lemmas. Write

$\boldsymbol{\beta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$ where $\boldsymbol{\alpha}$ is a $q_n \times 1$ vector and $\boldsymbol{\gamma}$ is a $(p_n - q_n) \times 1$ vector. Analogously, write $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\boldsymbol{\alpha}}^{(k)\top}, \hat{\boldsymbol{\gamma}}^{(k)\top})^\top$ and

$$g(\boldsymbol{\beta}) = \{\mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{D}(\boldsymbol{\beta})\}^{-1} \mathbf{X}^\top \mathbf{y} = \{\boldsymbol{\alpha}^*(\boldsymbol{\beta})^\top, \boldsymbol{\gamma}^*(\boldsymbol{\beta})^\top\}^\top. \quad (\text{A.1})$$

For simplicity, we write $\boldsymbol{\alpha}^*(\boldsymbol{\beta})$ and $\boldsymbol{\gamma}^*(\boldsymbol{\beta})$ as $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ hereafter. We further partition $\boldsymbol{\Sigma}_n^{-1}$ as

$$\boldsymbol{\Sigma}_n^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\top & \mathbf{A}_{22} \end{pmatrix}$$

where the $\mathbf{A}_{11}$ is a $q_n \times q_n$ matrix. Then, multiplying $(\mathbf{X}^\top \mathbf{X})^{-1} \{\mathbf{X}^\top \mathbf{X} + \lambda_n \mathbf{D}(\boldsymbol{\beta})\}$ to Eq. (A.1) gives

$$\begin{pmatrix} \boldsymbol{\alpha}^* - \boldsymbol{\beta}_{01} \\ \boldsymbol{\gamma}^* \end{pmatrix} + \frac{\lambda_n}{n} \begin{pmatrix} \mathbf{A}_{11} \mathbf{D}_1(\boldsymbol{\alpha}) \boldsymbol{\alpha}^* + \mathbf{A}_{12} \mathbf{D}_2(\boldsymbol{\gamma}) \boldsymbol{\gamma}^* \\ \mathbf{A}_{12}^\top \mathbf{D}_1(\boldsymbol{\alpha}) \boldsymbol{\alpha}^* + \mathbf{A}_{22} \mathbf{D}_2(\boldsymbol{\gamma}) \boldsymbol{\gamma}^* \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}, \quad (\text{A.2})$$

where $\mathbf{D}_1(\boldsymbol{\alpha}) = \text{diag}(\alpha_1^{-2}, \dots, \alpha_{q_n}^{-2})$ and $\mathbf{D}_2(\boldsymbol{\gamma}) = \text{diag}(\gamma_1^{-2}, \dots, \gamma_{p_n - q_n}^{-2})$.

Lemma 1. Let δ_n be a sequence of positive real numbers satisfying $\delta_n \rightarrow \infty$ and

$p_n \delta_n^2 / \lambda_n \rightarrow 0$. Define $\mathbf{H}_n = \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta_n \sqrt{p_n / n}\}$ and

$\mathbf{H}_{n1} \equiv \{\boldsymbol{\alpha} \in \mathbb{R}^{q_n} : \|\boldsymbol{\alpha} - \boldsymbol{\beta}_{01}\| \leq \delta_n \sqrt{p_n / n}\}$. Assume conditions (C1)–(C3) hold. Then, with probability tending to 1, we have

- (a) $\sup_{\boldsymbol{\beta} \in \mathbf{H}_n} \|\boldsymbol{\gamma}^*\| / \|\boldsymbol{\gamma}\| < 1/C_0$ for some constant $C_0 > 1$;
- (b) g is a mapping from $\mathbf{H}_n$ to itself.

Proof. We first prove part (a). Because $\lambda_n/\sqrt{n} \rightarrow 0$ and $p_n \delta_n^2/\lambda_n \rightarrow 0$, we have $\delta_n \sqrt{p_n/n} \rightarrow 0$.

By (C2), $nE(\|\hat{\beta}(OLS) - \beta_0\|^2) = nE\left(\left\|(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}\right\|^2\right) = \sigma^2 \text{trace}(\boldsymbol{\Sigma}_n^{-1}) = O(p_n)$ Hence,

$\|\hat{\beta}(OLS) - \beta_0\|^2 = O_p(p_n/n)$. It then follows from (A.2) that

$$\sup_{\beta \in \mathbf{H}_n} \left\| \gamma^* + \lambda_n \mathbf{A}_{12}^\top \mathbf{D}_1(\alpha) \alpha^*/n + \lambda_n \mathbf{A}_{22} \mathbf{D}_2(\gamma) \gamma^*/n \right\| = O_p(\sqrt{p_n/n}). \quad (\text{A.3})$$

Note that $\|\alpha - \beta_{01}\| \leq \delta_n(p_n/n)^{1/2}$ and $\|\alpha^*\| \leq \|g(\beta)\| \leq \|\hat{\beta}(OLS)\| = O_p(a_{1n} \sqrt{p_n})$ By assumptions (C2) and (C3), we have

$$\begin{aligned} \sup_{\beta \in \mathbf{H}_n} \left\| \lambda_n \mathbf{A}_{12}^\top \mathbf{D}_1(\alpha) \alpha^*/n \right\| &\leq \frac{\lambda_n}{n} \left\| \mathbf{A}_{12}^\top \right\| \sup_{\beta \in \mathbf{H}_n} \left\| \mathbf{D}_1(\alpha) \alpha^* \right\| \leq \sqrt{2} C \frac{\lambda_n a_{1n}}{n a_{0n}^2} \sup_{\beta \in \mathbf{H}_n} \left\| \alpha^* \right\| \quad (\text{A.4}) \\ &= o_p(\sqrt{p_n/n}), \end{aligned}$$

where the second inequality uses the fact $\left\| \mathbf{A}_{12}^\top \right\| \leq \sqrt{2} C$, which follows from the inequality $\left\| \mathbf{A}_{12} \mathbf{A}_{12}^\top \right\| - \left\| \mathbf{A}_{11}^2 \right\| \leq \left\| \mathbf{A}_{11}^2 + \mathbf{A}_{12} \mathbf{A}_{21} \right\| \leq \left\| \boldsymbol{\Sigma}_n^{-2} \right\| < C^2$. Combining (A.3) and (A.4) gives

$$\sup_{\beta \in \mathbf{H}_n} \left\| \gamma^* + \lambda_n \mathbf{A}_{22} \mathbf{D}_2(\gamma) \gamma^*/n \right\| = O_p(\sqrt{p_n/n}). \quad (\text{A.5})$$

Note that $\mathbf{A}_{22}$ is positive definite and by the singular value decomposition,

$\mathbf{A}_{22} = \sum_{i=1}^{p_n - q_n} \tau_{2i} \mathbf{u}_{2i} \mathbf{u}_{2i}^\top$ where τ_{2i} and $\mathbf{u}_{2i}$ are eigenvalues and eigenvectors of $\mathbf{A}_{22}$. Then, since $1/C < \tau_{2i} < C$, we have

$$\begin{aligned} \frac{\lambda_n}{n} \left\| \mathbf{A}_{22} \mathbf{D}_2(\gamma) \gamma^* \right\| &= \frac{\lambda_n}{n} \left\| \sum_{i=1}^{p_n - q_n} \tau_{2i} \mathbf{u}_{2i} \mathbf{u}_{2i}^\top \mathbf{D}_2(\gamma) \gamma^* \right\| = \frac{\lambda_n}{n} \left\{ \sum_{i=1}^{p_n - q_n} \tau_{2i}^2 \left\| \mathbf{u}_{2i}^\top \mathbf{D}_2(\gamma) \gamma^* \right\|^2 \right\}^{1/2} \\ &\geq \frac{\lambda_n}{n C} \left\{ \sum_{i=1}^{p_n - q_n} \left\| \mathbf{u}_{2i}^\top \mathbf{D}_2(\gamma) \gamma^* \right\|^2 \right\}^{1/2} = \frac{1}{C} \left\| \lambda_n \mathbf{D}_2(\gamma) \gamma^*/n \right\|. \end{aligned}$$

This, together with (A.5) and (C2), implies that with probability tending to 1,

$$\frac{1}{C} \left\| \lambda_n \mathbf{D}_2(\gamma) \gamma^*/n \right\| - \left\| \gamma^* \right\| \leq \delta_n \sqrt{p_n/n}. \quad (\text{A.6})$$

Let $\mathbf{d}_{\gamma^*/\gamma} = (\gamma_1^*/\gamma_1, \dots, \gamma_{p_n - q_n}^*/\gamma_{p_n - q_n})^\top$. Because $\|\gamma\| \leq \delta_n \sqrt{p_n/n}$ we have

$$\frac{1}{C} \left\| \frac{\lambda_n}{n} \mathbf{D}_2(\gamma) \gamma^* \right\| = \frac{1}{C} \frac{\lambda_n}{n} \left\| \{\mathbf{D}_2(\gamma)\}^{1/2} \mathbf{d}_{\gamma^*/\gamma} \right\| \geq \frac{1}{C} \frac{\lambda_n}{n} \frac{\sqrt{n}}{\delta_n \sqrt{p_n}} \left\| \mathbf{d}_{\gamma^*/\gamma} \right\| \quad (\text{A.7})$$

and

$$\|\gamma^*\| = \left\| \mathbf{D}_2(\gamma)^{-1/2} \mathbf{d}_{\gamma^*/\gamma} \right\| \leq \frac{\delta_n \sqrt{p_n}}{\sqrt{n}} \left\| \mathbf{d}_{\gamma^*/\gamma} \right\|. \quad (\text{A.8})$$

Combining (A.6), (A.7) and (A.8), we have that with probability tending to 1,

$$\left\| \mathbf{d}_{\gamma^*/\gamma} \right\| \leq \frac{1}{\lambda_n / (p_n \delta_n^2 C) - 1} < 1/C_0 \quad (\text{A.9})$$

for some constant $C_0 > 1$ provided that $\lambda_n / (p_n \delta_n^2) \rightarrow \infty$. It is worth noting that $\Pr(\left\| \mathbf{d}_{\gamma^*/\gamma} \right\| \rightarrow 0) \rightarrow 1$, as $n \rightarrow \infty$.

Furthermore, with probability tending to 1,

$$\|\gamma^*\| \leq \left\| \mathbf{d}_{\gamma^*/\gamma} \right\|_{1 \leq j \leq \max(p_n - q_n)} \left| \gamma_j \right| \leq \left\| \mathbf{d}_{\gamma^*/\gamma} \right\| \times \|\gamma\| \leq \|\gamma\| / C_0.$$

This proves part (a).

Next we prove part (b). First, it is easy to see from (A.8) and (A.9) that, as $n \rightarrow \infty$,

$$\Pr(\|\gamma^*\| \leq \delta_n \sqrt{p_n/n}) \rightarrow 1. \quad (\text{A.10})$$

Then, by (A.2), we have

$$\sup_{\beta \in \mathbf{H}_n} \left\| \alpha^* - \beta_{01} + \lambda_n \mathbf{A}_{11} \mathbf{D}_1(\alpha) \alpha^* / n + \lambda_n \mathbf{A}_{12} \mathbf{D}_2(\gamma) \gamma^* / n \right\| = O_p(\sqrt{p_n/n}). \quad (\text{A.11})$$

Similar to (A.4), it is easily to verify that

$$\sup_{\beta \in \mathbf{H}_n} \left\| \lambda_n \mathbf{A}_{11} \mathbf{D}_1(\alpha) \alpha^* / n \right\| = o_p(\sqrt{p_n/n}). \quad (\text{A.12})$$

Moreover, with probability tending to 1,

$$\sup_{\beta \in \mathbf{H}_n} \|\lambda_n \mathbf{A}_{12} \mathbf{D}_2(\gamma) \gamma^* / n\| \leq \frac{\lambda_n}{n} \sup_{\beta \in \mathbf{H}_n} \|\mathbf{D}_2(\gamma) \gamma^*\| \times \|\mathbf{A}_{12}\| \leq 2\sqrt{2}C^2 \delta_n \sqrt{p_n/n}, \quad (\text{A.13})$$

where the last step follows from (A.6), (A.10), and the fact that $\|\mathbf{A}_{12}\| \leq \sqrt{2}C$. It follows from (A.11), (A.12) and (A.13) that with probability tending to 1,

$$\sup_{\beta \in \mathbf{H}_n} \|\alpha^* - \beta_{01}\| \leq (2\sqrt{2}C^2 + 1) \delta_n n^{-1/2} \sqrt{p_n}. \quad (\text{A.14})$$

Because $\delta_n \sqrt{p_n}/\sqrt{n} \rightarrow 0$ we have, as $n \rightarrow \infty$,

$$\Pr(\alpha^* \in \mathbf{H}_{n1}) \rightarrow 1. \quad (\text{A.15})$$

Combining (A.10) and (A.15) completes the proof of part (b).

Lemma 2. Assume that (C1)–(C3) hold. For any q_n -vector $\mathbf{b}_n$ satisfying $\|\mathbf{b}_n\| \leq 1$, define $s_n^2 = \sigma^2 \mathbf{b}_n^\top \Sigma_{n1}^{-1} \mathbf{b}_n$ as in Theorem 1. Define

$$f(\alpha) = \{\mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\alpha)\}^{-1} \mathbf{X}_1^\top \mathbf{y}. \quad (\text{A.16})$$

Then, with probability tending to 1,

- (a) $f(\alpha)$ is a contraction mapping from $\mathbf{B}_n \equiv \{\alpha \in \mathbb{R}^{q_n} : \|\alpha - \beta_{01}\| \leq \delta_n \sqrt{p_n/n}\}$ to itself;
- (b) $\sqrt{n} s_n^{-1} \mathbf{b}_n^\top (\hat{\alpha}^\circ - \beta_{01}) \rightsquigarrow \mathcal{N}(0, 1)$ where $\hat{\alpha}^\circ$ is the unique fixed point of $f(\alpha)$ defined by

$$\hat{\alpha}^\circ = \{\mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\hat{\alpha}^\circ)\}^{-1} \mathbf{X}_1^\top \mathbf{y}.$$

Proof. We first prove part (a). Note that (A.16) can be rewritten as

$$f(\alpha) - \beta_{01} + \lambda_n \Sigma_{n1}^{-1} \mathbf{D}_1(\alpha) f(\alpha) / n = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \epsilon.$$

Thus,

$$\sup_{\alpha \in \mathbf{B}_n} \|f(\alpha) - \beta_{01} + \lambda_n \Sigma_{n1}^{-1} \mathbf{D}_1(\alpha) f(\alpha)/n\| = O_p(\sqrt{q_n/n}). \quad (\text{A.17})$$

Similar to (A.4), it can be shown that

$$\sup_{\alpha \in \mathbf{B}_n} \|\lambda_n \Sigma_{n1}^{-1} \mathbf{D}_1(\alpha) f(\alpha)/n\| = o_p(\sqrt{q_n/n}). \quad (\text{A.18})$$

It follows from (A.17) and (A.18) that

$$\sup_{\alpha \in \mathbf{B}_n} \|f(\alpha) - \beta_{01}\| \leq \delta_n \sqrt{q_n/n},$$

where δ_n is a sequence of real numbers satisfying $\delta_n \rightarrow \infty$ and $\delta_n \sqrt{q_n/n} \rightarrow 0$. This implies that, as $n \rightarrow \infty$,

$$\Pr\{f(\alpha) \in \mathbf{B}_n\} \rightarrow 1. \quad (\text{A.19})$$

In other words, f is a mapping from the region $\mathbf{B}_n$ to itself.

Rewrite (A.16) as $\{\mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\alpha)\} f(\alpha) = \mathbf{X}_1^\top \mathbf{y}$ and then differentiate it with respect to α , we have

$$\{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\alpha)/n\} \dot{f}(\alpha) + (\lambda_n/n) \times \text{diag}\{-2f_j(\alpha)/\alpha_j^3\} = 0,$$

where $\dot{f}(\alpha) = \partial f(\alpha)/\partial \alpha^\top$ and $\text{diag}\{-2f_j(\alpha)/\alpha_j^3\} = \text{diag}\{-2f_1(\alpha)/\alpha_1^3, \dots, -2f_{q_n}(\alpha)/\alpha_{q_n}^3\}$. This, together with the assumption $\lambda_n/\sqrt{n} \rightarrow 0$, implies that

$$\sup_{\alpha \in \mathbf{B}_n} \|\{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\alpha)/n\} \dot{f}(\alpha)\| = \sup_{\alpha \in \mathbf{B}_n} \frac{2\lambda_n}{n} \|\text{diag}\{f_j(\alpha)/\alpha_j^3\}\| = o_p(1). \quad (\text{A.20})$$

Note that Σ_{n1} is positive definite. Write $\Sigma_{n1} = \sum_{i=1}^{q_n} \tau_{1i} \mathbf{u}_{1i} \mathbf{u}_{1i}^\top$, where τ_{1i} and $\mathbf{u}_{1i}$ are eigenvalues and eigenvectors of Σ_{n1} . Then, by (C2), $\tau_{1i} \in (1/C, C)$ for all i and

$$\begin{aligned}
\|\Sigma_{n1}\dot{f}(\alpha)\| &= \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbb{R}^{q_n}} \|\Sigma_{n1}\dot{f}(\alpha)\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbb{R}^{q_n}} \left\| \sum_{i=1}^{q_n} \lambda_{1i} \mathbf{u}_{1i} \mathbf{u}_{1i}^\top \dot{f}(\alpha) \mathbf{x} \right\| \quad (\text{A.21}) \\
&= \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbb{R}^{q_n}} \left\{ \sum_{i=1}^{q_n} \lambda_{1i}^2 \|\mathbf{u}_{1i}^\top \dot{f}(\alpha) \mathbf{x}\|^2 \right\}^{1/2} \geq \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbb{R}^{q_n}} \frac{1}{C} \left\{ \sum_{i=1}^{q_n} \|\mathbf{u}_{1i}^\top \dot{f}(\alpha) \mathbf{x}\|^2 \right\}^{1/2} \\
&= \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbb{R}^{q_n}} \frac{1}{C} \|\dot{f}(\alpha)\mathbf{x}\| = \frac{1}{C} \|\dot{f}(\alpha)\|.
\end{aligned}$$

Therefore, it follows from $\alpha \in \mathbf{B}_n$ (A.21) and (C2) that

$$\|\{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\alpha)/n\}\dot{f}(\alpha)\| \geq \|\Sigma_{n1}\dot{f}(\alpha)\| - \|\lambda_n \mathbf{D}_1(\alpha)\dot{f}(\alpha)/n\| \geq \frac{1}{C} \|\dot{f}(\alpha)\| - \frac{\lambda_n}{n} a_{0n}^{-2} \|\dot{f}(\alpha)\|,$$

This, together with (A.20) and (C2), implies that

$$\sup_{\alpha \in \mathbf{B}_n} \|\dot{f}(\alpha)\| = o_p(1). \quad (\text{A.22})$$

Finally, the conclusion in part (a) follows from (A.19) and (A.22).

Next we prove part (b). Write

$$\begin{aligned}
n^{1/2} s_n^{-1} \mathbf{b}_n^\top (\hat{\alpha}^\circ - \beta_{01}) &= n^{1/2} s_n^{-1} \mathbf{b}_n^\top \left[\{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\hat{\alpha}^\circ)/n\}^{-1} \Sigma_{n1} - \mathbf{I}_{q_n} \beta_{01} \right] \quad (\text{A.23}) \\
&+ n^{1/2} s_n^{-1} \mathbf{b}_n^\top \{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\hat{\alpha}^\circ)/n\}^{-1} \mathbf{X}_1^\top \boldsymbol{\varepsilon} \equiv I_1 + I_2.
\end{aligned}$$

By the first order resolvent expansion formula, viz. $(\mathbf{H} + \lambda \mathbf{D})^{-1} = \mathbf{H}^{-1} - \lambda \mathbf{H}^{-1} \mathbf{D} \mathbf{H}^{-1} + o_p(\lambda)$, the first term on the right-hand side of Eq. (A.23) can be written as

$$I_1 = -s_n^{-1} \mathbf{b}_n^\top \Sigma_{n1}^{-1} \frac{\lambda_n}{\sqrt{n}} \mathbf{D}_1(\hat{\alpha}^\circ) \{\Sigma_{n1} + \lambda_n \mathbf{D}_1(\hat{\alpha}^\circ)/n\}^{-1} \Sigma_{n1} \beta_{01}.$$

Hence, by the assumptions (C2) and (C3), we have

$$\|I_1\| \leq \frac{\lambda_n}{\sqrt{n}} s_n^{-1} a_{0n}^{-2} \|\Sigma_{n1}^{-1} \beta_{01}\| = O_p\left(\lambda_n \frac{a_{1n}}{a_{0n}^2} \sqrt{\frac{q_n}{n}}\right) \rightarrow 0. \quad (\text{A.24})$$

Furthermore, by writing $X_1^\top = (\tilde{w}_1, \dots, \tilde{w}_n)$ and applying the first order resolvent expansion formula, it can be shown that

$$I_2 = \frac{s_n^{-1}}{\sqrt{n}} \sum_{i=1}^n \mathbf{b}_n^\top \Sigma_n^{-1} \tilde{\mathbf{w}}_i \varepsilon_i + o_p(1), \quad (\text{A.25})$$

which converges in distribution to $N(0, 1)$ by the Lindeberg–Feller Central Limit Theorem. Finally, combining (A.23), (A.24), and (A.25), we can conclude the proof of part (b).

Proof of Theorem 1. Note that for the initial ridge estimator $\hat{\beta}^{(0)}$ defined by (1), we have

$$\hat{\beta}^{(0)} - \beta_0 = \left(\Sigma_n + \xi_n \mathbf{I}_{p_n/n} \right)^{-1} \Sigma_n^{-1} \beta_0 + \left(\Sigma_n + \xi_n \mathbf{I}_{p_n/n} \right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} / n \equiv \mathbf{T}_1 + \mathbf{T}_2.$$

By the first order resolvent expansion formula and $\xi_n / \sqrt{n} \rightarrow 0$,

$$\|\mathbf{T}_1\| = \left\| -\Sigma_n^{-1} \xi_n \left(\Sigma_n + \xi_n \mathbf{I}_{p_n/n} \right)^{-1} \Sigma_n \beta_0 / n \right\| \leq C^3 \xi_n a_{1n} \sqrt{p_n/n} = o_p(\sqrt{p_n/n}).$$

It is easy to see that $\|\mathbf{T}_2\| = O_p(\sqrt{p_n/n})$. Thus $\|\hat{\beta}^{(0)} - \beta_0\| = O_p\{(p_n/n)^{1/2}\}$. This, combined with part (a) of Lemma 1, implies that

$$\Pr\left(\lim_{k \rightarrow \infty} \hat{\gamma}^{(k)} = 0\right) \rightarrow 1. \quad (\text{A.26})$$

Hence, to prove part (i) of Theorem 1, it is sufficient to show that

$$\Pr\left(\lim_{k \rightarrow \infty} \|\hat{\alpha}^{(k)} - \hat{\alpha}^*\| = 0\right) \rightarrow 1, \quad (\text{A.27})$$

where $\hat{\alpha}^*$ is the fixed point of $f(\mathbf{a})$ defined in part (b) of Lemma 2.

Define $\gamma^* = 0$ if $\gamma \neq 0$. It is easy to see from (A.2) that for any $\mathbf{a} \in \mathbf{B}_n$,

$$\lim_{\gamma \rightarrow 0} \gamma^*(\mathbf{a}, \gamma) = 0. \quad (\text{A.28})$$

Combining (A.28) with the fact

$$\begin{pmatrix} \mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\alpha) & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 + \lambda_n \mathbf{D}_2(\gamma) \end{pmatrix} \begin{pmatrix} \alpha^* \\ \gamma^* \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1^\top \mathbf{y} \\ \mathbf{X}_2^\top \mathbf{y} \end{pmatrix},$$

we find that, for any $\alpha \in \mathbf{B}_n$,

$$\lim_{\gamma \rightarrow 0} \alpha^*(\alpha, \gamma) = \{\mathbf{X}_1^\top \mathbf{X}_1 + \lambda_n \mathbf{D}_1(\alpha)\}^{-1} \mathbf{X}_1^\top \mathbf{y} = f(\alpha). \quad (\text{A.29})$$

Therefore, g is continuous and thus uniformly continuous on the compact set $\beta \in \mathbf{H}_n$. This, together with (A.26) and (A.29), implies that, as $k \rightarrow \infty$,

$$\eta_k \equiv \sup_{\alpha \in \mathbf{B}_n} \|f(\alpha) - \alpha^*(\alpha, \hat{\gamma}^{(k)})\| \rightarrow 0 \quad (\text{A.30})$$

with probability tending to 1. Note that

$$\|\hat{\alpha}^{(k+1)} - \hat{\alpha}^\circ\| = \|\alpha^*(\hat{\beta}^{(k)}) - \hat{\alpha}^\circ\| \leq \|\alpha^*(\hat{\beta}^{(k)}) - f(\hat{\alpha}^{(k)})\| + \|f(\hat{\alpha}^{(k)}) - \hat{\alpha}^\circ\| \leq \eta_k + \frac{1}{C} \|\hat{\alpha}^{(k)} - \hat{\alpha}^\circ\|,$$

where the last step follows from $\|f(\hat{\alpha}^{(k)}) - \hat{\alpha}^\circ\| = \|f(\hat{\alpha}^{(k)}) - f(\hat{\alpha}^\circ)\| \leq \|\hat{\alpha}^{(k)} - \hat{\alpha}^\circ\|/C$. Let $a_k = \|\hat{\alpha}^{(k)} - \hat{\alpha}^\circ\|$, for every integer $k \geq 0$. From (A.30), we can induce that with probability tending to 1, for any $\epsilon > 0$, there exists a positive integer N such that for every integer $k > N$, $|\eta_k| < \epsilon$ and

$$\begin{aligned} a_{k+1} &\leq \frac{a_k - 1}{C^2} + \frac{\eta_k - 1}{C} + \eta_k \leq \frac{a_1}{C^k} + \frac{\eta_1}{C^{k-1}} + \cdots + \frac{\eta_N}{C^{k-N}} + \left(\frac{\eta_{N+1}}{C^{k-N-1}} + \cdots + \frac{\eta_k - 1}{C} + \eta_k \right) \\ &\leq (a_1 + \eta_1 + \cdots + \eta_N) \frac{1}{C^{k-N}} + \frac{1 - (1/C)^{k-N}}{1 - 1/C} \in \end{aligned}$$

and the right-hand term tends to 0 as $k \rightarrow \infty$. This proves (A.27).

Therefore, it follows immediately from (A.26) and (A.27) that with probability tending to 1, $\lim_{k \rightarrow \infty} \beta^{(k)} = \lim_{k \rightarrow \infty} (\hat{\alpha}^{(k)\top}, \hat{\gamma}^{(k)\top})^\top = (\hat{\alpha}^{\circ\top}, 0)^\top$, which completes the proof of part (i).

This, in addition to part (b) of Lemma 2, proves part (ii) of Theorem 1

Proof of Theorem 2. Recall that $\hat{\beta}^* = \lim_{k \rightarrow \infty} \hat{\beta}^{(k+1)}$ and $\hat{\beta}^{(k+1)} = \arg \min_{\beta} \{Q(\beta | \hat{\beta}^{(k)})\}$,

where

$$Q(\beta | \hat{\beta}^{(k)}) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_n \sum_{\ell=1}^{p_n} \beta_\ell^2 \left\{ \hat{\beta}_\ell^{(k)} \right\}^2.$$

If $\beta_\ell^* \neq 0$ for $\ell \in \{i, j\}$, then $\hat{\beta}^*$ must satisfy the following normal equations for $\ell \in \{i, j\}$

$$-2\mathbf{x}_\ell^\top \{\mathbf{y} - \mathbf{X}\hat{\beta}^{(k+1)}\} + 2\lambda_n \hat{\beta}_\ell^{(k+1)} / \left\{ \hat{\beta}_\ell^{(k)} \right\}^2 = 0.$$

Thus, for $\ell \in \{i, j\}$,

$$\hat{\beta}_\ell^{(k+1)} / \left\{ \hat{\beta}_\ell^{(k)} \right\}^2 = \mathbf{x}_\ell^\top \hat{\mathbf{e}}^{(k+1)} / \lambda_n, \quad (\text{A.31})$$

where $\hat{\mathbf{e}}^{(k+1)} = \mathbf{y} - \mathbf{X}\hat{\beta}^{(k+1)}$. Moreover, because

$$\left\| \hat{\mathbf{e}}^{(k+1)} \right\|^2 + \lambda_n \sum_{i=1}^{p_n} \frac{\hat{\beta}_i^2}{\hat{\beta}_i^2} = Q(\hat{\beta}^{(k+1)} | \hat{\beta}^{(k)}) \leq Q(0 | \hat{\beta}^{(k)}) = \|\mathbf{y}\|^2,$$

we have

$$\left\| \hat{\mathbf{e}}^{(k+1)} \right\| \leq \|\mathbf{y}\| \quad (\text{A.32})$$

Letting $k \rightarrow \infty$ in (A.31) and (A.32), we have, for $\ell \in \{i, j\}$, and

$$\|\hat{\mathbf{e}}\| \leq \|\mathbf{y}\|, \beta_\ell^{*-1} = \mathbf{x}_\ell^\top \hat{\mathbf{e}} \lambda_n \text{ where } \hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}^*.$$

Therefore,

$$\left| \hat{\beta}_i^{*-1} - \hat{\beta}_j^{*-1} \right| \leq \frac{1}{\lambda_n} \|\mathbf{y}\| \times \|\mathbf{x}_i - \mathbf{x}_j\| \frac{1}{\lambda_n} \|\mathbf{y}\| \sqrt{2(1 - \rho_{ij})}.$$

and the proof is complete.

Proof of Theorem 3. (a) It is easy to check that

$$\begin{aligned}
\ell\{\boldsymbol{\beta}^{(t)}\} &= h\{\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t)}\} \geq h\{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}\} = \ell\{\boldsymbol{\beta}^{(t)}\} + \{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\}^\top S\{\boldsymbol{\beta}^{(t)}\} + \frac{u}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\
&= \ell\{\boldsymbol{\beta}^{(t+1)}\} + \frac{u}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + \left[-\ell\{\boldsymbol{\beta}^{(t+1)}\} + \ell\{\boldsymbol{\beta}^{(t)}\} + \{\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\}^\top S\{\boldsymbol{\beta}^{(t)}\} \right] \\
&= \ell\{\boldsymbol{\beta}^{(t+1)}\} + \frac{u}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 - \|\mathbf{X}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})\|_2^2 \\
&\geq \ell\{\boldsymbol{\beta}^{(t+1)}\} + \frac{u}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 - \lambda_{\max}(\mathbf{X}^\top \mathbf{X}) \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\
&\geq \ell\{\boldsymbol{\beta}^{(t+1)}\}
\end{aligned}$$

(A.33)

where the last step follows from the assumption that $u \geq 2\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$.

(b) The proof for the convergence of $\{\boldsymbol{\beta}^{(t)}\}$ to a local solution of (6) essentially mimicks the proof of Theorem 1 of Xu and Chen (2014) and thus is omitted here.

Proof of Theorem 4. Define $S_+^k = \{s : s_0 \subset s; \|s\|_0 \leq k\}$ and $S_-^k = \{s : s_0 \not\subset s; \|s\|_0 \leq k\}$ to be the collections the over-fitted models and the under-fitted models. Let $\hat{\boldsymbol{\beta}}_s$ be the estimate of $\boldsymbol{\beta}$ based on the model s . Clearly the result holds if $\Pr(\hat{s}_{\text{SLE}} \in S_+^k) \rightarrow 1$. Thus, it suffices to show that as $n \rightarrow \infty$,

$$\Pr \left\{ \min_{s \in S_-^k} \ell(\hat{\boldsymbol{\beta}}_s) \leq \max_{s \in S_+^k} \ell(\hat{\boldsymbol{\beta}}_s) \right\} \rightarrow 0.$$

For any $s \in S_-^k$, $s' = s \cup s_0 \in S_+^{2k}$. Consider $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s_0}\|_2 \omega_1 n^{-\tau_1}$, and by the Taylor expansion and conditions (D1) and (D3), we have

$$\begin{aligned}
\ell(\boldsymbol{\beta}_{s_0}) - \ell(\boldsymbol{\beta}_{s'}) &= (\boldsymbol{\beta}_{s_0} - \boldsymbol{\beta}_{s'})^\top S(\boldsymbol{\beta}_{s_0}) - \frac{1}{2} (\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s_0})^\top 2\mathbf{X}_{s'}^\top \mathbf{X}_{s'} (\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s_0}) \\
&\leq (\boldsymbol{\beta}_{s_0} - \boldsymbol{\beta}_{s'})^\top S(\boldsymbol{\beta}_{s_0}) - c_1 n \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s_0}\|_2^2 \leq \omega_1 n^{-\tau_1} \|S(\boldsymbol{\beta}_{s_0})\|_2 - c_1 \omega_1^2 n^{1-2\tau_1}.
\end{aligned}$$

Thus, upon setting $c = \omega_2^{-1/2} c_1 \omega_1 / \sqrt{2}$, we have

$$\begin{aligned}
\Pr\{\ell(\boldsymbol{\beta}_{s'0}) - \ell(\boldsymbol{\beta}_{s'}) \geq 0\} &\leq \Pr\left\{\|S(\boldsymbol{\beta}_{s'0})\|_2 \geq c_1 \omega_1 n^{1-\tau_1}\right\} \\
&\leq \sum_{j \in s'} \Pr\left\{S_j^2(\boldsymbol{\beta}_{s'0}) \geq \frac{1}{2k}(c_1 \omega_1)^2 n^{2-2\tau_1}\right\} = \sum_{j \in s'} \Pr\left\{|S_j(\boldsymbol{\beta}_{s'0})| \geq \frac{1}{(2k)^{1/2}} c_1 \omega_1 n^{1-\tau_1}\right\} \\
&\leq \sum_{j \in s'} \Pr\left\{|S_j(\boldsymbol{\beta}_{s'0})| \geq c n^{1-\tau_1-\tau_2/2}\right\}.
\end{aligned}$$

(A.34)

Let $t_{ni} = x_{ij} \left(\sum_{i=1}^n x_{ij}^2 \sigma_i^2 \right)^{-1/2}$. Such that $\sum_{i=1}^n t_{ni}^2 \sigma_i^2 = 1$. Under condition (D4), we have $\max_i(t_{ni}^2) \leq c_4/n$ and $\sum_{i=1}^n x_{ij}^2 \sigma_i^2 \leq c_2^2 n \sigma^2$ for some positive constant σ . Then,

$$\begin{aligned}
\Pr\left\{S_j(\boldsymbol{\beta}_{s'0}) \leq -c n^{1-\tau_1-\tau_2/2}/2\right\} &= \Pr\left\{2 \sum_{i=1}^n x_{ij} (y_i - \mathbf{w}_{is'}^\top \boldsymbol{\beta}_{s'0}) \geq c n^{1-\tau_1-\tau_2/2}/2\right\} \\
&\leq \Pr\left\{2 \sum_{i=1}^n t_{ni} (y_i - \mathbf{w}_{is'}^\top \boldsymbol{\beta}_{s'0}) \geq \frac{c}{2c_2 \sigma} n^{1/2-\tau_1-\tau_2/2}\right\} \leq \exp\left(-c' n^{1-2\tau_1-\tau_2}\right),
\end{aligned}$$

where the last step uses Lemma A.1 in [46] and c' is a positive constant. Similarly,

$$\Pr\left\{S_j(\boldsymbol{\beta}_{s'0}) \geq c n^{1-\tau_1-\tau_2/2}\right\} \leq \exp\left(-c' n^{1-2\tau_1-\tau_2}\right). \quad (\text{A.35})$$

Combining (A.34)–(A.35), we have $\Pr\{\ell(\boldsymbol{\beta}_{s'0}) - \ell(\boldsymbol{\beta}_{s'}) \geq 0\} \leq 4k \exp\left(-c' n^{1-2\tau_1-\tau_2}\right)$.

Hence,

$$\begin{aligned}
\Pr\left\{\min_{s \in S_-^k} \ell(\boldsymbol{\beta}_{s'}) \leq \ell(\boldsymbol{\beta}_{s'0})\right\} &\leq \sum_{s \in S_-^k} \Pr\{\ell(\boldsymbol{\beta}_{s'}) \leq \ell(\boldsymbol{\beta}_{s'0})\} \\
&\leq 4k p_n^k \exp\left(-c' n^{1-2\tau_1-\tau_2}\right) = 4 \exp\left(\ln k + k \ln p_n - c' n^{1-2\tau_1-\tau_2}\right) \\
&\leq 4w_2 \exp\left\{\tau_2 \ln n + O\left(n^{\tau_2+d}\right) - c' n^{1-2\tau_1-\tau_2}\right\} = o(1),
\end{aligned}$$

where the last two steps follow from conditions (D1)–(D2). Because $\ell(\boldsymbol{\beta}_{s'})$ is convex $\boldsymbol{\beta}_{s'}$, the above result holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'0}\|_2 \geq \omega_1 n^{-\tau_1}$.

For any $s \in S_-^k$, let $\check{\beta}_{s'}$ be $\hat{\beta}_s$ augmented with zeros corresponding to the elements in s'/s_0 . By condition (D1), we have $\|\check{\beta}_{s'} - \beta_{s'_0}\|_2 \geq \omega_1 n^{-\tau_1}$. Thus

$$\Pr \left\{ \min_{s \in S_-^k} \ell(\hat{\beta}_s) \leq \max_{s \in S_+^k} \ell(\hat{\beta}_s) \right\} \leq \Pr \left\{ \min_{s \in S_-^k} \ell(\check{\beta}_{s'}) \leq \ell(\beta_{s'_0}) \right\} = o(1).$$

This concludes the argument.

References

- [1]. Akaike H, A new look at the statistical model identification, IEEE Trans. Automat. Contr 19 (1974) 716–723.
- [2]. Andoh A, Nishida A, Takahashi K, Inatomi O, Imaeda H, Bamba S, Kito K, Sugimoto M, Kobayashi T, Comparison of the gut microbial community between obese and lean peoples using 16s gene sequencing in a Japanese population, J. Clin. Biochem. Nutr 59 (2016) 65–70. [PubMed: 27499582]
- [3]. Basso N, Soricelli E, Castagneto-Gissey L, Casella G, Albanese D, Fava F, Donati C, Tuohy K, Angelini G, La Neve F, Severino A, Kamvissi-Lorenz V, Birkenfeld AL, Bornstein S, Manco M, Mingrone G, Insulin resistance, microbiota, and fat distribution changes by a new model of vertical sleeve gastrectomy in obese rats, Diabetes 65 (2016) 2990–3001. [PubMed: 27431457]
- [4]. Blumensath T, Davies ME, Iterative hard thresholding for compressed sensing, Appl. Comput. Harmon. Anal 27 (2009) 265–274.
- [5]. Breheny P, Huang J, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, Ann. Appl. Statist 5 (2011) 232–253.
- [6]. Breiman L, Heuristics of instability and stabilization in model selection, Ann. Statist 24 (1996) 2350–2383.
- [7]. Candès EJ, Wakin MB, Boyd SP. Enhancing sparsity by reweighted ℓ_1 minimization, J. Fourier Anal. Appl 14 (2008) 877–905.
- [8]. Chartrand R, Yin W, Iteratively reweighted algorithms for compressive sensing, In: Proceedings of Int. Conf. on Acoustics, Speech, Signal Processing (ICASSP) (2008) 3869–3872.
- [9]. Chen J, Chen Z, Extended bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.
- [10]. Daubechies I, DeVore R, Fornasier M, Güntürk CS, Iteratively reweighted least squares minimization for sparse recovery, Commun. Pure Appl. Math 63 (2010) 1–38.
- [11]. Dicker L, Huang B, Lin X, Variable selection and estimation with the seamless- ℓ_0 penalty, Stat. Sinica 23 (2013) 929–962.
- [12]. Fan J, Feng Y, Song R, Nonparametric independence screening in sparse ultra-high-dimensional additive models, J. Amer. Statist. Assoc 106 (2011) 544–557.
- [13]. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc 96 (2001) 1348–1360.
- [14]. Fan J, Lv J, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol 70 (2008) 849–911.
- [15]. Fan J, Xue L, Zou H, Strong oracle optimality of folded concave penalized estimation, Ann. Statist 42 (2014) 819–849.
- [16]. Federico A, Dallio M, Tolone S, Gravina AG, Patrone V, Romano M, Tuccillo C, Mozzillo AL, Amoroso V, Misso G, Morelli L, Docimo L, Loguercio C, Gastrointestinal hormones, intestinal microbiota and metabolic homeostasis in obese patients: Effect of bariatric surgery, in vivo 30 (2016) 321–330. [PubMed: 27107092]

- [17]. Foster DP, George EI, The risk inflation criterion for multiple regression, *Ann. Statist* 22 (1994) 1947–1975.
- [18]. Friedman J, Hastie T, Tibshirani RJ, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw* 33 (2010) 1. [PubMed: 20808728]
- [19]. Frommlet F, Nuel G, An adaptive ridge procedure for ℓ_1 regularization, *PloS one* 11 (2016) e0148620. [PubMed: 26849123]
- [20]. Gasso G, Rakotomamonjy A, Canu S, Recovering sparse signals with a certain family of nonconvex penalties and DC programming, *IEEE Trans. Signal Process* 57 (2009) 4686–4698.
- [21]. Gorodnitsky IF, Rao BD, Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm, *IEEE Trans. Signal Process* 45 (1997) 600–616.
- [22]. He X, Wang L, Hong HG, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *Ann. Statist* 41 (2013) 342–369.
- [23]. Hippe B, Remely M, Aumueller E, Pointner A, Magnet U, Haslberger A, *Faecalibacterium prausnitzii* phylotypes in type two diabetic, obese, and lean control subjects, *Benef. Microbes* 7 (2016) 511–517. [PubMed: 27048834]
- [24]. Huang J, Horowitz JL, Ma S, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Ann. Statist* 36 (2008) 587–613.
- [25]. Johnson BA, Long Q, Huang Y, Chansky K, Redman M, Log-penalized least squares, iteratively reweighted lasso, and variable selection for censored lifetime medical cost, Department of Biostatistics and Bioinformatics, Technical Report, Emory University, Atlanta, GA, 2012.
- [26]. Jung M-J, Lee J, Shin N-R, Kim M-S, Hyun D-W, Yun J-H, Kim PS, Whon TW, Bae J-W, Chronic repression of mTOR complex 2 induces changes in the gut microbiota of diet-induced obese mice, *Sci. Rep* 6 (2016) 30887. [PubMed: 27471110]
- [27]. Knight K, Fu W, Asymptotics for lasso-type estimators, *Ann. Statist* 28 (2000) 1356–1378.
- [28]. Lai P, Liu Y, Liu Z, Wan Y, Model free feature screening for ultrahigh dimensional data with responses missing at random, *Comput. Statist. Data Anal* 105 (2017) 201–216.
- [29]. Lawson CL, Contribution to the theory of linear least maximum approximation, Doctoral dissertation, Univ. Calif. at Los Angeles, Los Angeles, CA, 1961.
- [30]. Li G, Peng H, Zhang J, Zhu L, Robust rank correlation based screening, *Ann. Statist* 40 (2012) 1846–1877.
- [31]. Lin L, Sun J, Adaptive conditional feature screening, *Comput. Statist. Data Anal* 94 (2016) 287–301.
- [32]. Liu J, Li R, Wu R, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, *J. Amer. Statist. Assoc* 109 (2014) 266–274.
- [33]. Liu Z, Li G, Efficient regularized regression with penalty for variable selection and network construction, *Comput. Math. Methods Med* 2016 (2016) ID 3456153.
- [34]. Loh P-L, Wainwright MJ, Regularized M -estimators with nonconvexity: Statistical and algorithmic theory for local optima, *J. Mach. Learn. Res* 16 (2015) 559–616.
- [35]. Mallows CL, Some comments on C_p , *Technometrics* 15 (1973) 661–675.
- [36]. Meinshausen N, Bühlmann P, High-dimensional graphs and variable selection with the lasso, *Ann. Statist* 34 (2006) 1436–1462.
- [37]. Osborne MR, *Finite algorithms in optimization and data analysis*, Wiley, 1985.
- [38]. Schwarz G, Estimating the dimension of a model, *Ann. Statist* 6 (1978) 461–464.
- [39]. Segal MR, Dahlquist KD, Conklin BR, Regression approaches for microarray data analysis, *J. Comput. Biol* 10 (2003) 961–980. [PubMed: 14980020]
- [40]. Shen X, Pan W, Zhu Y, Likelihood-based selection and sharp parameter estimation, *J. Amer. Statist. Assoc* 107 (2012) 223–232.
- [41]. Shen X, Pan W, Zhu Y, Zhou H, On constrained and regularized high-dimensional regression, *Ann. Inst. Statist. Math* 65 (2013) 807–832.
- [42]. Stamey TA, Kabalin JN, McNeal JE, Johnstone IM, Freiha F, Redwine EA, Yang N, Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II: Radical prostatectomy treated patients, *J. Urol* 141 (1989) 1076–1083. [PubMed: 2468795]

- [43]. Tibshirani RJ, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series B Stat. Methodol 58 (1996) 267–288.
- [44]. Wang H, Forward regression for ultra-high dimensional variable screening, J. Amer. Statist. Assoc 104 (2009) 1512–1524.
- [45]. Wipf D, Nagarajan S, Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions, IEEE J. Sel. Topics Signal Process 4 (2) (2010) 317–329.
- [46]. Xu C, Chen J, The sparse mle for ultrahigh-dimensional feature screening, J. Amer. Statist. Assoc 109 (2014) 1257–1269.
- [47]. Zhang C-H, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist 38 (2010) 894–942.
- [48]. Zhao P, Yu B, On model selection consistency of lasso, J. Mach. Learn. Res 7 (2006) 2541–2563.
- [49]. Zhong W, Zhu L, Li R, Cui H, Regularized quantile regression and robust feature screening for single index models, Stat. Sinica 26 (2016) 69–95.
- [50]. Zhu L-P, Li L, Li R, Zhu L-X, Model-free feature screening for ultrahigh-dimensional data, J. Amer. Statist. Assoc 106 (2011) 1464–1475.
- [51]. Zou H, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc 101 (2006) 1418–1429.
- [52]. Zou H, Hastie T, Regularization and variable selection via the Elastic Net, J. R. Stat. Soc. Series B Stat. Methodol 67 (2005) 301–320.
- [53]. Zupancic ML, Cantarel BL, Liu Z, Drabek EF, Ryan KA, Cirimotich S, Jones C, Knight R, Walters WA, Knights D, Mongodin EF, Horenstein RB, Mitchell BD, Steinle N, Snitker S, Shuldiner AR, Fraser CM, Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome, PloS one 7 (2012) e43052. [PubMed: 22905200]

Figure 1:
Thresholding functions of BAR, L_0 , Lasso, Adaptive Lasso, SCAD, and $L_{0.5}$.

Figure 2:

Path plots of BAR, Elastic Net, Lasso, SCAD, Adaptive Lasso, and MCP for a linear model with two groups of highly correlated covariates with true coefficients $\beta_j = 2$ for $j=1,2$, and 3; $\beta_j = 0.5$ for $j = 4, 5$, and 6; and $\beta_j = 0$ for $j = 7$, as described in Section 3.2. For BAR, we fixed $\xi_n = 100$. For Elastic Net, the fixed tuning parameter is $\lambda_2 = 0.001$ and SCAD with $\gamma = 3.7$, Adaptive Lasso with $\xi = 18.68$, MCP with $\gamma = 3$. The dashed vertical lines are the optimal tuning parameters obtained by 5-fold CV with lowest CV error and the corresponding estimated coefficients are provided in Table 2.

Table 1:

(Low-dimensional settings: $n = 100$, $p \in \{10, 50, 100\}$ Comparison of BAR, Lasso, SCAD, MCP, and Adaptive Lasso (ALasso) and TLP based on 100 Monte-Carlo samples. (Misclassification = mean number of misclassified non-zeros and zeros; FP = mean number of false positives (non-zeros); FN = mean number of false negatives (zeros); TM = probability that the selected model is exactly the true model; MAB = mean absolute bias; MSPE = mean squared prediction error from five-fold CV.)

Model	p	Method	Misclassification	FP	FN	TM	MAB	MSPE
1	10	BAR	0.08	0.08	0	93 %	0.26	1.02
		Lasso	3.28	3.28	0	6 %	0.56	1.08
		SCAD	0.90	0.90	0	62 %	0.34	1.03
		MCP	0.72	0.72	0	68 %	0.33	1.03
		ALasso	0.26	0.26	0	89 %	0.29	1.03
		TLP(0.15)	0.54	0.54	0	63 %	0.34	1.04
		TLP(0.5)	0.64	0.64	0	77 %	0.31	1.04
	50	BAR	0.26	0.26	0	86 %	0.32	1.04
		Lasso	7.75	7.75	0	2 %	0.95	1.18
		SCAD	1.81	1.81	0	65 %	0.37	1.05
		MCP	0.97	0.97	0	73 %	0.34	1.05
		ALasso	3.99	3.99	0	40 %	0.68	1.03
		TLP(0.15)	0.98	0.98	0	60 %	0.45	1.05
		TLP(0.5)	1.51	1.51	0	77 %	0.36	1.08
	100	BAR	0.44	0.44	0	80 %	0.35	1.02
		Lasso	10.38	10.38	0	2 %	1.08	1.2
		SCAD	1.41	1.41	0	65 %	0.31	1.03
		MCP	1.04	1.04	0	66 %	0.32	1.03
		ALasso	1.51	1.51	0	32 %	0.42	1.08
		TLP(0.15)	1.31	1.31	0	55 %	0.46	1.02
		TLP(0.5)	2.21	2.21	0	70 %	0.38	1.05
2	10	BAR	1.14	0.12	1.02	30 %	0.67	1.11
		Lasso	2.70	2.61	0.09	5 %	0.76	1.11
		SCAD	1.87	1.58	0.29	14 %	0.74	1.1
		MCP	1.69	1.27	0.42	19 %	0.74	1.1
		ALasso	1.55	1.11	0.44	18 %	0.69	1.07
		TLP(0.15)	1.02	0.40	0.62	33 %	0.66	1.07
		TLP(0.5)	2.27	1.87	0.4	9 %	0.74	1.09
	50	BAR	2.18	0.68	1.5	4 %	0.94	1.17
		Lasso	10.56	10.09	0.47	0 %	1.51	1.29
		SCAD	5.36	4.32	1.04	1 %	1.06	1.19
		MCP	3.33	2.1	1.23	4 %	0.99	1.19
		ALasso	8.22	7.38	0.84	0 %	1.41	1.1
		TLP(0.15)	2.25	1.36	1.19	7 %	1.03	1.1
		TLP(0.5)	6.07	5.13	0.94	3 %	1.09	1.2

Model	p	Method	Misclassification	FP	FN	TM	MAB	MSPE
	100	BAR	2.67	0.94	1.73	2%	1.05	1.17
		Lasso	14.63	14.02	0.61	0%	1.79	1.34
		SCAD	4.24	3.08	1.16	1%	1.01	1.21
		MCP	3.34	1.99	1.35	3%	0.96	1.2
		ALasso	4.42	2.27	2.15	1%	1.26	1.32
		TLP(0.15)	3.14	1.88	1.26	2%	1.03	1.08
		TLP(0.5)	7.67	6.71	0.96	1%	1.17	1.19

Table 2:

(Grouping effects) Estimated coefficients of BAR, Elastic Net, Lasso, SCAD, Adaptive Lasso, and MCP for a linear model with optimal tuning parameters selected by five-fold CV. The full solution paths are depicted in Figure 2.

Parameter	BAR	Elastic Net	Lasso	SCAD	Adaptive Lasso	MCP
(Intercept)	-0.008	-0.004	-0.005	0.031	-0.009	0.029
β_1	1.902	2.000	2.137	6.001	1.585	6.001
β_2	2.016	1.998	1.441		1.908	
β_3	2.081	1.999	2.421		2.506	
β_4	0.499	0.406	0.894	1.200	1.153	1.200
β_5	0.469	0.401	0.292		0.047	
β_6	0.233	0.392	0.013			
β_7				-0.001		
β_8						
β_9						
Tuning	$\xi_n = 100$	$\lambda_1 = 0.017$	$\lambda = 0.017$	$\gamma = 3.7,$	$\xi = 18.68$	$\gamma = 3,$
parameters	$\lambda_n = 0.043$	$\lambda_2 = 0.001$		$\lambda = 0.044$	$\lambda = 0.001$	$\lambda = 10.861$
CV Error	0.038	0.045	0.331	0.048	0.043	0.330

Table 3:

(High-dimensional settings: $n = 100$, $p = 2000$) Comparison of BAR, Lasso, SCAD, MCP, Adaptive Lasso (ALasso), and TLP Following the Sparsity-Restricted Least Squares Estimate (sLSE). (Misclassification = Mean Number of Misclassified Non-zeros and Zeros; FP = Mean Number of False Positives (Non-zeros); FN = Mean Number of False Negatives (Zeros); TM = Probability that the Selected Model is Exactly the True Model; MAB = Mean Absolute Bias; MSPE = Mean Squared Prediction Error from Five-Fold CV.)

Model	Method	Misclassification	FP	FN	TM	MAB	MSPE
1	sLSE-BAR	0.32	0.18	0.14	75.72 %	0.55	1.04
	sLSE-Lasso	5.10	4.96	0.14	3.8 %	0.97	1.13
	sLSE-SCAD	0.94	0.80	0.14	64.60 %	0.57	1.05
	sLSE-MCP	0.67	0.53	0.14	68.32 %	0.57	1.05
	sLSE-ALasso	0.92	0.78	0.14	63.88 %	0.61	1.05
	sLSE-TLP(0.15)	0.74	0.60	0.14	56.56 %	0.64	1.06
	sLSE-TLP(0.5)	0.89	0.75	0.14	67.72 %	0.57	1.06
2	sLSE-BAR	0.30	0.18	0.12	0 %	1.45	1.04
	sLSE-Lasso	5.09	4.98	0.11	0 %	1.87	1.13
	sLSE-SCAD	0.95	0.83	0.12	0 %	1.47	1.05
	sLSE-MCP	0.66	0.54	0.12	0 %	1.47	1.05
	sLSE-ALasso	0.93	0.81	0.12	0 %	1.51	1.05
	sLSE-TLP(0.15)	0.72	0.60	0.12	0 %	1.53	1.06
	sLSE-TLP(0.5)	0.91	0.79	0.12	0 %	1.47	1.07

Table 4:

Coefficients of selected features and test mean squared prediction error for a prostate data (with $p = 36$ main effects and two-way interactions)

Variable	BAR	Lasso	SCAD	MCP	Adaptive Lasso	Elastic Net
lcavol	0.573	0.560	0.637	0.627	0.594	0.264
lweight					0.020	0.056
Age		-0.028	-0.073	-0.149	-0.077	-0.053
lbph						0.070
svi		0.125	0.422	0.440	0.141	0.090
lcp						0.001
gleason						0.013
(lcavol)×(lweight)						0.068
(lcavol)×(age)						0.036
(lcavol)×(svi)	0.341	0.216			0.233	0.150
(lcavol)×(gleason)						0.196
(lweight)×(gleason)	0.315	0.238	0.243	0.305	0.259	0.171
(age)×(lbph)		0.184	0.277	0.282	0.194	0.098
(age)×(svi)						0.045
(lbph)×(gleason)						0.021
(svi)×(gleason)		0.030				0.090
(lcp)×(pgg45)			-0.019	-0.067		
Tuning	$\xi_n = 44.45$	$\lambda = 0.08$	$\gamma = 3.7$	$\gamma = 3$	$\xi = 1.16$	$\lambda_1 = 0.06$
Parameters	$\lambda_n = 0.1$		$\lambda = 0.08$	$\lambda = 0.07$	$\lambda = 0.21$	$\lambda_2 = 0.1$
Number of Selected Features	3	7	6	6	7	16
Test Error	0.461	0.484	0.587	0.583	0.513	0.498

Table 5:

Coefficients of selected BMI-related features and test mean squared prediction error for a gut microbiota metagenomic data

Taxon	BAR	Lasso	SCAD	MCP	Adaptive Lasso	Elastic Net
<i>Acetanaerobacterium</i>						-0.064
<i>Acidaminococcus</i>						-0.014
<i>Akkermansia</i>						0.233
<i>Anaerotruncus</i>						-0.105
<i>Asteroleplasma</i>						-0.058
<i>Bacteroides</i>						-0.110
<i>Barnesiella</i>						0.195
<i>Butyricicoccus</i>						-0.046
<i>Butyrivibrio</i>						0.205
<i>Catenibacterium</i>						0.198
<i>Clostridium</i>						-0.259
<i>Collinsella</i>	1.066	0.669	0.689	0.867	0.908	0.708
<i>Coprobacillus</i>		-0.165	-0.133	-0.064	-0.216	-0.295
<i>Coprococcus</i>						0.014
<i>Desulfovibrio</i>		-0.082	-0.049		-0.122	-0.308
<i>Escherichia Shigella</i>						0.081
<i>Eubacterium</i>						0.021
<i>Faecalibacterium</i>	-0.610	-0.389	-0.352	-0.352	-0.584	-0.591
<i>Hydrogenoanaerobacterium</i>		0.099	0.044		0.223	0.409
<i>Lachnobacterium</i>						0.294
<i>Lactobacillus</i>						0.115
<i>Megasphaera</i>	0.761	0.458	0.410	0.470	0.716	0.639
<i>Odoribacter</i>						-0.024
<i>Parabacteroides</i>						-0.043
<i>Prevotella</i>						0.335
<i>Rikenella</i>						-0.065
<i>Robinsoniella</i>		-0.113	-0.088	-0.050	-0.124	-0.344
<i>Sporobacter</i>						-0.102
<i>Streptococcus</i>						0.050
<i>Subdoligranulum</i>						-0.018
<i>Succiniclasicum</i>		0.119	0.080		0.211	0.344
<i>Succinivibrio</i>						0.238
<i>Sutterella</i>		-0.001				-0.224
<i>Treponema</i>					-0.115	
<i>Turicibacter</i>	-0.996	-0.622	-0.615	-0.807	-0.846	-0.611
<i>Xylanibacter</i>						0.160
Tuning	$\xi_n = 77.78$	$\lambda = 0.54$	$\gamma = 3.7$	$\gamma = 3$	$\xi = 14.19$	$\lambda_1 = 0.19$
Parameters	$\lambda_n = 38$		$\lambda = 0.57$	$\lambda = 0.66$	$\lambda = 3.67$	$\lambda_2 = 0.26$

Taxon	BAR	Lasso	SCAD	MCP	Adaptive Lasso	Elastic Net
Number of Selected Taxa	4	10	9	6	9	36
Test Error	28.630	28.189	28.311	28.230	27.926	27.062

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript