Statistica Sinica **28** (2018), 2633-2655
doi:https://doi.org/10.5705/ss.202017.0027

PARTIAL CONSISTENCY WITH SPARSE INCIDENTAL PARAMETERS

Jianqing Fan[1], Runlong Tang[2] and Xiaofeng Shi[1]

[1]*Princeton University and* [2]*Johns Hopkins University*

Abstract: The penalized estimation principle is fundamental to high-dimensional problems. In the literature, it has been extensively and successfully applied to various models with only structural parameters. In this paper, we apply this penalization principle to a linear regression model with not only structural parameters but also sparse incidental parameters. For the estimators of the structural parameters, we derive their consistency and asymptotic normality, which reveals an oracle property. However, the penalized estimators for the incidental parameters possess only partial selection consistency, not consistency. This is an interesting partial consistency phenomenon: the structural parameters are consistently estimated while the incidental ones are not. For the structural parameters, also considered is an alternative two-step penalized estimator, which has fewer possible asymptotic distributions and thus is more suitable for statistical inferences. A data-driven approach for selecting a penalty regularization parameter is provided. The finite-sample performance of the penalized estimators for the structural parameters is evaluated by simulations and a data set is analyzed. We also extend the methods and results to the case where the number of the structural parameters diverge but slower than the sample size.

Key words and phrases: Oracle property, partial consistency, penalized estimation, sparse incidental parameter, structural parameter, two-step estimation.

1. Introduction

Since the pioneering papers by Tibshirani (1996) and Fan and Li (2001), the penalized estimation methodology exploiting sparsity has been studied extensively. For example, Zhao and Yu (2006) provides an almost necessary and sufficient condition, the Irrepresentable Condition, for the LASSO estimator to be strong sign consistent. Fan, Liao and Mincheva (2011) shows that an oracle property holds for the folded concave penalized estimator with ultrahigh dimensionality. For an overview on this topic, see Fan and Lv (2010).

These papers consider models only with structural parameters that are related to every data point. Here we consider another type of model where there

are not only the structural parameters but also incidental parameters, each of which is related to only one data point. Specifically, suppose data $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ are from the linear model

$$Y_i = \mu_i^\star + \boldsymbol{X}_i^T \boldsymbol{\beta}^\star + \epsilon_i, \tag{1.1}$$

where the vector of incidental parameters $\boldsymbol{\mu}^\star = (\mu_1^\star, \ldots, \mu_n^\star)^T$ is sparse, the vector of structural parameters $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_d^\star)^T$ is of main interest, and $\boldsymbol{X}_i$'s are ϵ_i's are covariate vectors random errors, respectively. Let $\boldsymbol{\nu} = (\boldsymbol{\mu}^{\star T}, \boldsymbol{\beta}^{\star T})^T$. Then, a different data point $(\boldsymbol{X}_i, Y_i)$ depends on a different subset of $\boldsymbol{\nu}$, $\mu_i^\star$ and $\boldsymbol{\beta}^\star$.

Model (1.1) arises as a working model for estimation from Fan, Feng and Tong (2012), which considers a large-scale hypothesis testing problem under arbitrary dependence of test statistics. By principal factor approximation, a method proposed by Fan, Feng and Tong (2012), the dependent test statistics $\mathbf{Z} = (Z_1, \ldots, Z_p)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be decomposed as $Z_i = \mu_i + \boldsymbol{b}_i^T \boldsymbol{W} + K_i$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, $\boldsymbol{b}_i$ is the ith row of the first k unstandardized principal components, denoted by $\boldsymbol{B}$, of $\boldsymbol{\Sigma}$, and $\boldsymbol{K} = (K_1, \ldots, K_p)^T \sim N(0, \boldsymbol{A})$ with $\boldsymbol{A} = \boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{B}^T$. The common factor $\boldsymbol{W}$ drives the dependence among the test statistics. This realized but unobserved factor is critical for false discovery proportion (FDP) estimation and power improvements by removing the common factor $\{\boldsymbol{b}_i^T \boldsymbol{W}\}$ from the test statistics. Hence, an important goal is to estimate $\boldsymbol{W}$ with given $\{\boldsymbol{b}_i\}_{i=1}^n$. In many applications on large-scale hypothesis testing, $\{\mu_i\}_{i=1}^p$ are sparse. For example, genome-wide association studies show that the expression level of gene CCT8 is highly related to the phenotype of Down Syndrome. It is of interest to test the association between each of millions of SNP's and the CCT8 gene expression level. In the framework of Fan, Feng and Tong (2012), each μ_i stands for such an association: if $\mu_i = 0$, the ith SNP has no association with the CCT8 gene expression level; otherwise, it is associated. Since most of the SNP's are not associated the CCT8 gene expression level, it is reasonable to assume the sparsity of $\{\mu_i\}_{i=1}^p$. Replacing Z_i, μ_i, $\boldsymbol{b}_i$, $\boldsymbol{W}$, k, p, and K_i with Y_i, $\mu_i^\star$, $\boldsymbol{X}_i$, $\boldsymbol{\beta}^\star$, d, n, and ϵ_i, respectively, we obtain model (1.1).

Although model (1.1) emerges from a critical component of estimating FDP in Fan, Feng and Tong (2012), it has its own interest. For example, in some applications, those few nonzero $\mu_i^\star$'s might be some signals or measurement or recording errors of the responses $\{Y_i\}$ and what is interesting is to learn about $\boldsymbol{\beta}^\star$. Then, (1.1) is suitable for modeling data with "contaminated" responses and a method producing a reliable estimator for $\boldsymbol{\beta}^\star$ is a robust replacement for the ordinary least squares estimation that is known to be sensitive to outliers.

Several models with structural and incidental parameters were studied in a seminal paper by Neyman and Scott (1948), which points out the inconsistency of the maximum likelihood estimators (MLE) of structural parameters in the presence of a large number of incidental parameters and provides a modified MLE. Their method does not work for model (1.1) due to not exploiting the sparsity of incidental parameters. Kiefer and Wolfowitz (1956) shows the consistency of the MLE of the structural parameters when the incidental parameters are assumed to be from a common distribution; they basically eliminate the high-dimensionality issue of the incidental parameters by randomization. Our paper considers deterministic incidental parameters and handles the high-dimensionality issue by penalization with a sparsity assumption. Basu (1977) considers the elimination of nuisance parameters via marginalizing and conditioning methods, and Moreira (2009) solves the incidental parameter problem with an invariance principle. For a review of the incidental parameter problems in statistics and economics, see Lancaster (2000).

Without loss of generality, suppose the first s incidental parameters $\{\mu_i^\star\}_{i=1}^s$ are nonvanishing and the remainder are zero. Then, model (1.1) can be written in a matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\nu} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{X}_{1,s}^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_{s+1,n}^T & \boldsymbol{I}_{n-s} \end{pmatrix},$$

$\boldsymbol{X}_{i,j}^T = (\boldsymbol{X}_i, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_j)^T$, $\boldsymbol{I}_k$ is a $k \times k$ identity matrix, $\boldsymbol{0}$ is a generic block of zeros and $\boldsymbol{\nu} = (\mu_1^\star, \ldots, \mu_s^\star, \boldsymbol{\beta}^T, \mu_{s+1}^\star, \ldots, \mu_n^\star)^T$. Although this is a sparse high-dimensional problem, the matrix $\boldsymbol{X}$ does not satisfy the sufficient conditions for the results in Zhao and Yu (2006) and Fan, Liao and Mincheva (2011) due to the estimation inconsistency of the incidental parameters in $\boldsymbol{\nu}$. For details, see Supplement A.

In this paper, we propose a penalized estimator of $(\boldsymbol{\mu}^\star, \boldsymbol{\beta}^\star)$,

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}) = \underset{(\boldsymbol{\mu},\boldsymbol{\beta})\in\mathbb{R}^{n+d}}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \mu_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \sum_{i=1}^n p_\lambda(|\mu_i|), \tag{1.2}$$

where p_λ is a penalty function with a regularization parameter λ. The penalty is imposed only on the sparse incidental parameters. An iterative algorithm is proposed to compute the estimators. The estimator $\hat{\boldsymbol{\beta}}$ possesses consistency, asymptotic normality, and an oracle property. On the other hand, the nonvanishing elements of $\boldsymbol{\mu}^\star$ cannot be consistently estimated even if $\boldsymbol{\beta}^\star$ were known. So, there is a partial consistency phenomenon.

Penalized estimation (1.2) is a one-step method. For the estimation of $\boldsymbol{\beta}^\star$,

we also propose a two-step method whose first step is designed to eliminate the influence of the data with large incidental parameters. The two-step estimator $\tilde{\boldsymbol{\beta}}$ from the two-step method has fewer possible asymptotic distributions than $\hat{\boldsymbol{\beta}}$ and thus is more suitable for constructing confidence regions for $\boldsymbol{\beta}^\star$. The two-step estimator is asymptotically equivalent to the one-step estimator when the sizes of the nonzero incidental parameters are small enough. The two-step method improves the convergence rate and efficiency over the one-step method for challenging situations where large nonzero incidental parameters increase the asymptotic covariance or even reduce the convergence rate for the one-step method.

The rest of the paper is organized as follows. In Section 2, the model and penalized estimation method are rigorously introduced and the corresponding penalized estimators are characterized. In Section 3, asymptotic properties of the penalized estimators are derived, a penalized two-step estimator is proposed and its theoretical properties are obtained, and we provide a data-driven approach for selecting the regularization parameter. In Section 4, we present simulation results and analyze a data set. Section 5 concludes with a discussion. All the proofs and some additional theoretical results are relegated to an online supplementary file, where we also provide a study on the case where the number of covariates grows with, but slower than, the sample size.

2. Model and Method

The matrix form of model (1.1) is

$$\boldsymbol{Y} = \boldsymbol{\mu}^\star + \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. The covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) copies of $\boldsymbol{X}_0 \in \mathbb{R}^d$, a random vector with mean zero and a covariance matrix $\boldsymbol{\Sigma}_X > 0$. They are independent of the random errors $\{\epsilon_i\}$ that are i.i.d. copies of ϵ_0, a random variable with mean zero and variance $\sigma^2 > 0$. Write $a_n \ll b_n$ and $a_n \gg b_n$ if $a_n = o(b_n)$ and $b_n = o(a_n)$, respectively. There is an assumption on the covariates and random errors.

Assumption (A): There exist positive sequences $\kappa_n \ll \sqrt{n}$ and $\gamma_n \ll \sqrt{n}$ such that

$$P\left(\max_{1\le i\le n} \|\boldsymbol{X}_i\|_2 > \kappa_n\right) \to 0 \text{ and } P\left(\max_{1\le i\le n} |\epsilon_i| > \gamma_n\right) \to 0, \text{ as } n \to \infty, \tag{2.2}$$

where $\|\cdot\|_2$ stands for the l_2 norm of $\mathbb{R}^d$.

Suppose there are three types of incidental parameters: let $\{\mu_i^\star\}_{i=1}^{s_1}$ be large

in the sense that $|\mu_i^\star| \gg \max\{\kappa_n, \gamma_n\}$ for $1 \le i \le s_1$; $\{\mu_i^\star\}_{i=s_1+1}^{s}$ are nonzero and bounded by γ_n with $s = s_1 + s_2$; $\{\mu_i^\star\}_{i=s+1}^{n}$ are zero. It is unknown to us which $\mu_i^\star$'s are large, bounded, or zero. The sparsity of $\boldsymbol{\mu}^\star$ means $s_1 + s_2 \ll n$. Denote the vectors of the three types of incidental parameters as $\boldsymbol{\mu}_1^\star$, $\boldsymbol{\mu}_2^\star$, and $\boldsymbol{\mu}_3^\star$, respectively.

The penalized estimation (1.2) can be written as

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}) = \underset{(\boldsymbol{\mu},\boldsymbol{\beta})}{\operatorname{argmin}}\ L(\boldsymbol{\mu}, \boldsymbol{\beta}), \tag{2.3}$$

where $L(\boldsymbol{\mu}, \boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^n p_\lambda(|\mu_i|)$ and the penalty function p_λ can be the soft (i.e. L_1 or LASSO), hard, SCAD, or a general folded concave penalty function of Fan and Li (2001). For simplicity, here we consider only the soft penalty function $p_\lambda(|\mu_i|) = 2\lambda|\mu_i|$. The cases with the hard and SCAD penalties can be considered in a similar way.

By Lemma B.1 in Supplement B, a necessary and sufficient condition for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ to be a minimizer of $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ is that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, and $Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} = \lambda \mathrm{Sign}(\hat{\mu}_i)$ for $i \in \hat{I}_0^c$ and $|Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}| \le \lambda$ for $i \in \hat{I}_0$, where $\mathrm{Sign}(\cdot)$ is the sign function and $\hat{I}_0 = \{1 \le i \le n : \hat{\mu}_i = 0\}$.

The special structure of $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ suggests for the minimization problem in (2.3) a marginal decent algorithm that iteratively computes $\boldsymbol{\mu}^{(k)} = \operatorname{argmin}_{\boldsymbol{\mu}\in\mathbb{R}^n} L(\boldsymbol{\mu}, \boldsymbol{\beta}^{(k-1)})$ and $\boldsymbol{\beta}^{(k)} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} L(\boldsymbol{\mu}^{(k)}, \boldsymbol{\beta})$ until convergence. The advantage of this algorithm is that there exist analytic solutions to the two minimization steps. They are the soft-threshold estimators with residuals $\{Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}^{(k-1)}\}$ and the ordinary least-squares estimator with responses $\boldsymbol{Y} - \boldsymbol{\mu}^{(k)}$, respectively. A case with a diverging number of covariates d is considered in Supplement D. In this paper, d is assumed to be a fixed integer and we make an assumption on λ.

Assumption (B): The regularization parameter λ satisfies

$$\kappa_n \ll \lambda,\ \alpha\gamma_n \le \lambda,\ \text{and } \lambda \ll \min\{\mu^\star, \sqrt{n}\}, \tag{2.4}$$

where κ_n and γ_n are defined in (2.2), α is a constant greater than 2, and $\mu^\star = \min_{1\le i\le s_1} |\mu_i^\star|$.

Write "with probability going to one" as "wpg1". A stopping rule for the above algorithm is based on the successive difference $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2$. By Proposition B.1 in Supplement B, wpg1, the iterative algorithm stops at the the second iteration, given the initial estimator is bounded wpg1.

Suppose $\{\boldsymbol{\beta}^{(k)}\}$ has a theoretical limit $\boldsymbol{\beta}^{(\infty)}$, corresponding to which there is a limit estimator $\boldsymbol{\mu}^{(\infty)}$. Then, $(\boldsymbol{\mu}^{(\infty)}, \boldsymbol{\beta}^{(\infty)})$ is a solution of the equations

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}), \tag{2.5}$$

$$\boldsymbol{\mu} = (|\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta}| - \lambda)_+ \mathrm{Sign}(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta}), \tag{2.6}$$

where $(\cdot)_+$ returns the maximum value of the input and zero. By Lemma B.2 in Supplement B, a necessary and sufficient condition for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ to be a minimizer of $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ is that it is a solution to equations (2.5) and (2.6). Hence, $(\boldsymbol{\mu}^{(\infty)}, \boldsymbol{\beta}^{(\infty)})$ is a minimizer of $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ and can also be denoted as $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$.

The estimator $\hat{\boldsymbol{\beta}}$ is also the minimizer of the profiled loss function $\tilde{L}(\boldsymbol{\beta}) = L(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\beta})$, where $\boldsymbol{\mu}(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ is given by (2.6). Interestingly, this profiled loss function is a criterion function equipped with the famous Huber loss function (see Huber (1964, 1973)). Specifically, $\tilde{L}(\boldsymbol{\beta})$ can be expressed as $\sum_{i=1}^n \rho(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})$, where $\rho(x) = x^2 I(|x| \le \lambda) + (2\lambda x - \lambda^2) I(|x| > \lambda)$ is exactly the Huber's loss function, which is optimal in a minimax sense. This equivalence between the penalized estimation and Huber's robust estimation indicates that the penalization principle is versatile and can naturally produce an important loss function in robust statistics. It also provides a formal endorsement of the least absolute deviation robust regression (LAD) in Fan, Feng and Tong (2012) and indicates that it is better to use all data with LAD regression rather than 90% of them. The penalized estimation is only formally equal to the Huber's. Our model (2.1) considers deterministic sparse incidental parameters $\mu_i^\star$'s, while the model in Huber's works assumes random contamination as in Kiefer and Wolfowitz (1956). Recently, there are a few papers on robust regression in high-dimensional settings, see, for example, Chen, Wang and McKeown (2010), Lambert-Lacroix and Zwald (2011), Fan, Fan and Barut (2014), and Bean et al. (2012). Portnoy and He (2000) provides a review of literature on robust statistics.

From (2.5) and (2.6), $\hat{\boldsymbol{\beta}}$ is a solution to

$$\varphi_n(\boldsymbol{\beta}) = 0, \text{ with } \varphi_n(\boldsymbol{\beta}) = \boldsymbol{\beta} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}. \tag{2.7}$$

In general, this is a Z-estimation problem. The following analysis is based on this characterization of $\hat{\boldsymbol{\beta}}$.

At the end of this section, we provide some notations and an expansion of $\varphi_n(\boldsymbol{\beta})$. Let $\mathbb{S} = \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$, $\mathbb{S}_S = \sum_{i\in S} \boldsymbol{X}_i\boldsymbol{X}_i^T$, $\mathbb{S}_S^\mu = \sum_{i\in S} \boldsymbol{X}_i\mu_i^\star$, $\mathbb{S}_S^\epsilon = \sum_{i\in S} \boldsymbol{X}_i\epsilon_i$, $\mathcal{S} = \sum_{i=1}^n \boldsymbol{X}_i$, and $\mathcal{S}_S = \sum_{i\in S} \boldsymbol{X}_i$, where S is a subset of $\{1, \ldots, n\}$. It is straightforward to show

$$\begin{aligned}\varphi_n(\boldsymbol{\beta}) = {} & (\mathbb{S}_{S_{10}} + \mathbb{S}_{S_{11}} + \mathbb{S}_{S_{12}})(\boldsymbol{\beta} - \boldsymbol{\beta}^\star) - (\mathbb{S}^\mu_{S_{11}} + \mathbb{S}^\mu_{S_{12}}) - (\mathbb{S}^\epsilon_{S_{10}} + \mathbb{S}^\epsilon_{S_{11}} + \mathbb{S}^\epsilon_{S_{12}}) \\ & - \lambda(\mathcal{S}_{S_{20}} + \mathcal{S}_{S_{21}} + \mathcal{S}_{S_{22}} - \mathcal{S}_{S_{30}} - \mathcal{S}_{S_{31}} - \mathcal{S}_{S_{32}}),\end{aligned} \tag{2.8}$$

with the index sets $S_{10} = \{s+1 \le i \le n : |\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \boldsymbol{\beta}) + \epsilon_i| \le \lambda\}$, $S_{11} =$

$\{1 \le i \le s_1 : |\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \boldsymbol{\beta}) + \epsilon_i| \le \lambda\}$ and $S_{12} = \{s_1 + 1 \le i \le s : |\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \boldsymbol{\beta}) + \epsilon_i| \le \lambda\}$; S_{20}, S_{21} and S_{22} are defined similarly except that the absolute operation is omitted and "$\le$" is replaced by "$>$"; S_{30}, S_{31} and S_{32}, are defined similarly with S_{20}, S_{21} and S_{22} except that "$> \lambda$" is replaced by "$< -\lambda$". All these index sets depend on $\boldsymbol{\beta}$.

3. Asymptotic Properties

In this section, we consider the asymptotic properties of the penalized estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$. Assumptions (A) and (B) enable our method to distinguish the large incidental parameters from others, and thus simplify the asymptotic properties of the index sets S_{ij}'s in (2.8) by making them independent of $\boldsymbol{\beta}$ wpg1. Denote a hypercube of $\boldsymbol{\beta}^\star$ by $B_C(\boldsymbol{\beta}^\star) = \{\boldsymbol{\beta} \in \mathbb{R}^d : |\beta_j - \beta_j^\star| \le C, 1 \le j \le d\}$ with a constant $C > 0$.

Lemma 1 (On Index Sets S_{ij}'s)**.** *Under assumptions* (A) *and* (B)*, for every* $C > 0$ *and every* $\boldsymbol{\beta} \in B_C(\boldsymbol{\beta}^\star)$*, wpg1,* $S_{10} = S_{10}^\star$*,* $S_{11} = \emptyset$*,* $S_{12} = S_{12}^\star$*,* $S_{20} = \emptyset$*,* $S_{21} = S_{21}^\star$*,* $S_{22} = \emptyset$*,* $S_{30} = \emptyset$*,* $S_{31} = S_{31}^\star$ *and* $S_{32} = \emptyset$*, where the limit index sets* $S_{10}^\star = \{s+1, s+2, \ldots, n\}$*,* $S_{12}^\star = \{s_1+1, s+2, \ldots, s\}$*,* $S_{21}^\star = \{1 \le i \le s_1 : \mu_i^\star > 0\}$ *and* $S_{31}^\star = \{1 \le i \le s_1 : \mu_i^\star < 0\}$.

By Lemma 1, wpg1, the solution $\hat{\boldsymbol{\beta}}$ to (2.7) has an expression

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^\star + (\mathbb{S}_{S_{10}^\star} + \mathbb{S}_{S_{12}^\star})^{-1}\{\mathbb{S}^\mu_{S_{12}^\star} + (\mathbb{S}^\epsilon_{S_{10}^\star} + \mathbb{S}^\epsilon_{S_{12}^\star}) + \lambda(\mathcal{S}_{S_{21}^\star} - \mathcal{S}_{S_{31}^\star})\}, \tag{3.1}$$

from which we derive asymptotic properties of $\hat{\boldsymbol{\beta}}$. We need an assumption.

Assumption (C): There exists a constant $\delta > 0$ such that $\mathbb{E}\,\|\boldsymbol{X}_0\|_2^{2+\delta} < \infty$ and $\|\boldsymbol{\mu}_2^\star\|_2 / \|\boldsymbol{\mu}_2^\star\|_{2+\delta} \to \infty$, where $\|\boldsymbol{\mu}_2^\star\|_{2+\delta} = \left(\sum_{i=s_1+1}^{s} |\mu_i^\star|^{2+\delta}\right)^{1/(2+\delta)}$.

Theorem 1 (Existence and Consistency of $\hat{\boldsymbol{\beta}}$)**.** *Under assumptions* (A) *and* (B)*, if either* $s_2 = o(n/(\kappa_n\gamma_n))$ *or assumption* (C) *holds, then, for every fixed* $C > 0$*, wpg1, there exists a unique estimator* $\hat{\boldsymbol{\beta}}_n \in B_C(\boldsymbol{\beta}^\star)$ *such that* $\psi_n(\hat{\boldsymbol{\beta}}_n) = 0$ *and* $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}^\star$.

In Theorem 1, there are two kinds of sufficient conditions: one is on s_2, the size of bounded incidental parameters $\boldsymbol{\mu}_2^\star$, and the other is assumption (C), which is about the norms of $\boldsymbol{\mu}_2^\star$. They come from different analysis approaches to the term $\mathbb{S}^\mu_{S_{12}^\star}$ in (3.1). One does not imply the other. For details, see Supplement C. Specially, if $s_2 = O(n^{\alpha_2})$ for some $\alpha_2 \in (0, 1)$ and $\kappa_n\gamma_n \ll n^{(1-\alpha_2)}$, then $\hat{\boldsymbol{\beta}}$ is consistent by Theorem 1.

Next, we consider the asymptotic distributions of the consistent estimator

$\hat{\boldsymbol{\beta}}_n$ obtained in Theorem 1. Without loss of generality, we assume the sizes of index sets $S_{21}^\star = \{1 \le i \le s_1 : \mu_i^\star > 0\}$ and $S_{31}^\star = \{1 \le i \le s_1 : \mu_i^\star < 0\}$ are asymptotically equivalent to as_1 and $(1-a)s_1$, respectively, with a constant $a \in (0,1)$. Similar to Theorem 1, there are two sets of conditions on $\boldsymbol{\mu}_2^\star$ corresponding to two different analysis approaches. Denote $\sim$ as asymptotic equivalence and $D_n = \|\boldsymbol{\mu}_2^\star\|_2$.

Theorem 2 (Asymptotic Distributions on $\hat{\boldsymbol{\beta}}_n$)**.** *Under assumptions* (A) *and* (B)*, suppose either* $s_2 \ll \sqrt{n}/(\kappa_n\gamma_n)$ *holds or assumption* (C) *and* $D_n^2/n = o(1)$ *hold.*

(1) *If* $s_1 \ll n/\lambda^2$*, then* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$; [***main case***]

(2) *If* $s_1 \sim bn/\lambda^2$*, then* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (b+\sigma^2)\boldsymbol{\Sigma}_X^{-1})$*, for every constant* $b \in \mathbb{R}^+$*;*

(3) *If* $s_1 \gg n/\lambda^2$*, then* $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$*, where* $r_n \sim n/(\lambda\sqrt{s_1})$*.*

When the incidental parameters are sparse, the size s_1 of large incidental parameters is small. If the size s_2 or the magnitude D_n of bounded incidental parameters is also small, then the conclusion of case (1) tends to hold. This case is of most interest and we denote it as the main case. The other cases are presented to provide a relatively complete picture of the asymptotic distributions of $\hat{\boldsymbol{\beta}}$. In fact, Theorem C.1 in Supplement C shows more possible asymptotic distributions. The constant a does not appear in the limit distributions of Theorem 2 due to cancellation. The sub-$\sqrt{n}$ convergence rate emerges in case (3), because for this case the impact of the large incidental parameters is too big to be efficiently handled by the penalized estimation. For case (2), in one direction, as $b \to 0$, its condition and limit distribution become those of case (1); in the other direction, as b increases, it approaches case (3). This boundary phenomenon is in spirit similar to that in Tang, Banerjee and Kosorok (2012). Specially, if $\lambda \ll n^{\alpha_1}$, $\kappa_n\gamma_n \ll n^{\alpha_2}$, $s_1 \ll n^{1-\alpha_1}$, and $s_2 \ll n^{1/2-\alpha_2}$ for some $\alpha_1 \in (0,1)$ and $\alpha_2 \in (0,1/2)$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$ by the main case of Theorem 2.

Remark 1 (An Oracle Property). Suppose an oracle knows the true value of $\boldsymbol{\mu}^\star$. Then, with the adjusted responses $\boldsymbol{Y} - \boldsymbol{\mu}^\star$, the oracle estimator of $\boldsymbol{\beta}^\star$ is $\hat{\boldsymbol{\beta}}^{(O)} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}^\star)$. The limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{(O)} - \boldsymbol{\beta}^\star)$ is $N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$. Comparing this with the main case of Theorem 2, we see the penalized estimator $\hat{\boldsymbol{\beta}}_n$ enjoys an oracle property.

Although mainly interested in the estimation of $\boldsymbol{\beta}^\star$, we also obtain the soft-

threshold estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}^\star$: for each i,

$$\hat{\mu}_i = \mu_i(\hat{\boldsymbol{\beta}}) = (|Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}| - \lambda)_+ \operatorname{sgn}(Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}).$$

Let $\mathcal{E} = \{\hat{\mu}_i \neq 0, \text{ for } i = 1, \ldots, s_1, \text{ and } \hat{\mu}_i = 0, \text{ for } i = s_1 + 1, \ldots, n\}$.

Theorem 3 (Partial Selection Consistency on $\hat{\boldsymbol{\mu}}$)**.** *Under assumptions* (A) *and* (B)*, if* $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$*, then* $P(\mathcal{E}) \to 1$.

Theorem 3 shows that, wpg1, the indexes of $\boldsymbol{\mu}_{\boldsymbol{1}}^{\boldsymbol{\star}}$ and $\boldsymbol{\mu}_{\boldsymbol{3}}^{\boldsymbol{\star}}$ are estimated correctly, but those of $\boldsymbol{\mu}_{\boldsymbol{2}}^{\boldsymbol{\star}}$ wrongly. We call this a partial selection consistency phenomenon.

3.1. Two-step estimation

Theorem 2 tells us that the penalized estimator $\hat{\boldsymbol{\beta}}_n$ has multiple limit distributions, which complicate its application in practice. In addition, the convergence rate of $\hat{\boldsymbol{\beta}}_n$ is less than the optimal rate $\sqrt{n}$ in the case where the impact of large incidental parameters is substantial. To address these issues, we propose a two-step estimation method: first apply the penalized estimation (2.3) and let $\hat{I}_0 = \{1 \leq i \leq n : \hat{\mu}_i = 0\}$; then define the two-step estimator as $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{X}_{\hat{I}_0})^{-1}\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{Y}_{\hat{I}_0}$, where $\boldsymbol{X}_{\hat{I}_0}$ consists of $\boldsymbol{X}_i$'s whose indexes are in $\hat{I}_0$ and $\boldsymbol{Y}_{\hat{I}_0}$ consists of the corresponding Y_i's.

Theorem 4 (Consistency and Asymptotic Normality of $\tilde{\boldsymbol{\beta}}$)**.** *Suppose assumptions* (A) *and* (B) *hold. If either* $s_2 = o(n/(\kappa_n\gamma_n))$ *or assumption* (C) *holds, then* $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$*. If either* $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$ *holds or assumption* (C) *and* $D_n^2/n = o(1)$ *hold, then* $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$.

Comparing Theorem 4 with Theorem 2, we see that $\hat{\boldsymbol{\beta}}$ has three possible asymptotic distributions but $\tilde{\boldsymbol{\beta}}$ has only one since for $\tilde{\boldsymbol{\beta}}$ the conditions on s_1 disappear; this happens because the two-step method identifies and removes large incidental parameters by exploiting the partial selection consistency property of $\hat{\boldsymbol{\mu}}$ shown by Theorem 3. Further, the two-step estimator improves the convergence rate to the optimal one over the one-step estimator for the case with $s_1 \gg n/\lambda^2$. Due to these advantages, we suggest using the two-step method when making statistical inferences.

When the incidental parameters are sparse in the sense that $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$ or $D_n^2/n = o(1)$, it follows by Theorem 4 that $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$, from which a confidence region with asymptotic confidence level $1-\alpha$ is given by $\{\boldsymbol{\beta} \in \mathbb{R}^d : \sigma^{-1}\sqrt{n}\|\boldsymbol{\Sigma}_X^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \leq q_\alpha(\chi_d)\}$, where $q_\alpha(\chi_d)$ is the upper

α-quantile of χ_d, the square root of the chi-squared distribution with degrees of freedom d. For each component $\beta_j^\star$ of $\boldsymbol{\beta}^\star$, an asymptotic $1-\alpha$ confidence interval is

$$[\tilde{\beta}_j \pm n^{-1/2}\sigma \boldsymbol{\Sigma}_X^{-1/2}(j,j) z_{\alpha/2}], \tag{3.2}$$

where $\boldsymbol{\Sigma}_X^{-1/2}(j,j)$ is the square root of the (j,j) entry of $\boldsymbol{\Sigma}_X^{-1}$ and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$. The confidence region and interval involve unknown parameters $\boldsymbol{\Sigma}_X$ and σ. They can be estimated by $\hat{\boldsymbol{\Sigma}}_X = (1/n)\boldsymbol{X}^T\boldsymbol{X}$ and $\hat{\sigma} = \#(\hat{I}_0)^{-1/2}\|\boldsymbol{Y}_{\hat{I}_0} - \boldsymbol{X}_{\hat{I}_0}^T\tilde{\boldsymbol{\beta}}\|_2$, where $\#(\hat{I}_0)$ is the size of $\hat{I}_0$. By the law of large numbers, $\hat{\boldsymbol{\Sigma}}_X$ is consistent. By Lemma C.1 in Supplement C, $\hat{\sigma}$ is also consistent. Hence, after replacing $\boldsymbol{\Sigma}_X$ and σ in the confidence region and interval with $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\sigma}$, the resulting confidence region and interval have the asymptotic confidence level $1-\alpha$.

3.2. Theoretical and data-driven regularization parameters

Assumption (B) shows the theoretical regularization parameter λ depends on κ_n and γ_n, which are also crucial to the conditions of the asymptotic properties of the penalized estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Assumption (A) says κ_n and γ_n are determined by the distributions of $\boldsymbol{X}_0$ and ϵ_0, respectively. It is of interest to explicitly derive κ_n and γ_n for some typical cases of the covariates and errors. When the covariates are bounded with $C_X > 0$ and the random errors are $N(0,\sigma^2)$, let $\kappa_n = \sqrt{d}C_X$ and $\gamma_n = \sqrt{2\sigma^2\log(n)}$. They satisfy (2.2) in assumption (A), and the specification of λ (2.4) in assumption (B) becomes $\alpha\sqrt{2\sigma^2\log(n)} \le \lambda \ll \min\{\mu^\star, \sqrt{n}\}$. When $\boldsymbol{X}_0$ and ϵ_0 are $N(0,\boldsymbol{\Sigma}_X)$ and $N(0,\sigma^2)$, respectively. Denote by σ_X^2 the maximum of the diagonal elements of $\boldsymbol{\Sigma}_X$. Let $\kappa_n = \sqrt{2d\sigma_X^2\log(n)}$ and $\gamma_n = \sqrt{2\sigma^2\log(n)}$. They satisfy (2.2) in assumption (A), and (2.4) in assumption (C) becomes $\sqrt{\log(n)} \ll \lambda \ll \min\{\mu^\star, \sqrt{n}\}$. A case of exponentially tailed random variables is considered in Supplement C.

Although the theoretical specification of λ guaranties desired asymptotic properties, a data-driven specification is of interest. A popular way to specify λ is to use multi-fold cross-validation. The validation set, however, needs to be made as uncontaminated as possible. We propose a procedure to identify a data-driven λ:

Step 1: On the training and testing data sets.

1. Apply ordinary least squares (OLS) to all the data and obtain residuals $\hat{\epsilon}_i^{(OLS)} = Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}^{(OLS)}$ for each i.

2. Identify the set of "pure" data corresponding to the n_{pure} smallest values in $\{|\hat{\epsilon}_i^{(OLS)}|\}$.
3. Compute the updated OLS estimator $\hat{\boldsymbol{\beta}}^{(OLS,2)}$ with the "pure" data and obtain updated residuals $\{\hat{\epsilon}_i^{(OLS,2)}\}$ for each i.
4. Identify the updated "pure" data set corresponding to the n_{pure} smallest $\{|\hat{\epsilon}_i^{(OLS,2)}|\}$ and label the remaining as the "contaminated" data set.
5. Randomly select a subset from the updated "pure" data set as a testing set and merge the remaining "pure" data set and the "contaminated" one into a training set.

Step 2: On the range $[\lambda_L, \lambda_U]$ of the regularization parameter.

1. Compute the standard deviation $\hat{\sigma}_{pure}$ of the residuals of the "pure" data set.
2. Set $\lambda_L = \alpha_l \hat{\sigma}_{pure}$ and $\lambda_U = \alpha_u \hat{\sigma}_{pure}$, where $\alpha_l < \alpha_u$ are positive constants.

Step 3: On the data-driven regularization parameter.

1. For each grid point of λ in the interval $[\lambda_L, \lambda_U]$, apply a penalized method to the training set and obtain the estimator $\hat{\boldsymbol{\beta}}_{\lambda,train}$ and the corresponding test error $\hat{\sigma}^2_{\lambda,test} = \sum_{\text{testing set}} (Y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}_{\lambda,train})^2$.
2. Identify the data-driven regularization parameter λ_{opt} that minimizes $\hat{\sigma}^2_{\lambda,test}$, among the grid points.

This simple data-driven procedure can certainly be further improved. For example, In Step 1, the sub-steps 3 and 4 can be repeated more times to obtain a better "pure" data set. In Step two, the range for λ can also be obtained from quantiles of $\{|\hat{\epsilon}_i^{(OLS,2)}|\}$. We can also combine quantities based on $\hat{\sigma}_{pure}$ and quantiles of $\{|\hat{\epsilon}_i^{(OLS,2)}|\}$ to determine $[\lambda_L, \lambda_U]$.

The good performance of this data-driven regularization parameter are demonstrated in Subsection 4.2.

4. Numerical Evaluations and Data Analysis

We evaluated the finite-sample performance of the penalized estimation by simulations and applied it to a data set. The model for simulations was $Y_i =$

$\mu_i^\star + X_{i,1}\beta_1^\star + \cdots + X_{i,50}\beta_{50}^\star + \epsilon_i$ for $i = 1, \ldots, n$. The deterministic sparse incidental parameters $\{\mu_i^\star\}$ were generated as i.i.d. copies of μ: μ was 0, $U[-c, c]$, and $W(c + \mathrm{Exp}(\tau))$ with probabilities p_0, p_1, and p_2, respectively, and W took values 1 and -1 with probabilities p_w and $1 - p_w$, respectively, while $\mathrm{Exp}(\tau)$ was exponential with mean $1/\tau > 0$. Here c can be viewed as a contamination parameter: the larger c is, the more contaminated the data are. On the other hand, p_w determines the asymmetry of the incidental parameters. The regression coefficients were $\beta_1^\star = \cdots = \beta_{50}^\star = 1$; $\{(X_{i,1}, \ldots, X_{i,50})\} \overset{i.i.d.}{\sim} N(0, \mathbf{\Sigma}_X)$, where $\mathbf{\Sigma}_X(i, j) = 2\exp(-|i - j|)$, which is a Toeplitz matrix and the constant 2 was used to inflate the covariance; the covariates were independent of $\{\epsilon_i\} \overset{i.i.d.}{\sim} N(0, 1)$; $n = 500$; $p_0 = 0.8$, $p_1 = 0.1$, $p_2 = 0.1$, c was 0.5, 1, 3 or 5, p_w was 0.5 or 0.75, and $\tau = 1$.

4.1. Performance of penalized methods

The following methods for estimating $\boldsymbol{\beta}^\star$ were evaluated. (i) Oracle method (O): an oracle knows the index set S of zero $\mu_i^\star$'s; its performance is a benchmark. (ii) Ordinary least squares method (OLS): all $\mu_i^\star$'s were thought as zero. (iii) Four penalized least squares methods (PLS): PLS with soft penalty (PLS.Soft or S), PLS with hard penalty (PLS.Hard or H), two-step PLS with soft penalty (PLS.Soft.TwoStep or S.TS), and two-step PLS with hard penalty (PLS.Hard. TwoStep or H.TS). More specifically, the oracle estimator of $\boldsymbol{\beta}^\star$ is $\hat{\boldsymbol{\beta}}^{(O)} = (\sum_{i \in S} \boldsymbol{X}_i \boldsymbol{X}_i^T)^{-1} \sum_{i \in S} \boldsymbol{X}_i Y_i$; the hard penalty function is $p_\lambda(|t|) = \lambda^2 - (|t| - \lambda)^2\{|t| < \lambda\}$ (see Fan and Li (2001)). Each method was evaluated by the square root of the empirical mean squared error (RMSE). Every penalized method was implemented with some values of the regularization parameter λ, ranging from 0.5 to 5 by 0.25.

The sequence plot of Figure 1 shows 500 realized incidental parameters $\mu_i^\star$'s with $c = 3$ and $p_w = 0.75$, of which 94 are nonzero. They were used in the data generation of simulations. With fifty covariates, it is usually difficult to graphically identify the contaminated data points, as shown in the scatter plot of Figure 1.

Using the same incidental parameters, the six methods were evaluated by simulations with the iteration number 1,000. Since $\mathbf{\Sigma}_X$ is a Toeplitz matrix with equal diagonal elements, the asymptotic variances of the estimators of $\beta_1^\star$ and $\beta_2^\star$ are different, and representative for other $\beta_i^\star$'s. So, we only report results on the estimation of $\beta_1^\star$ and $\beta_2^\star$.

Figure 2 shows RMSE's of six estimators for $\beta_1^\star$. RMSE's for $\beta_2^\star$ are similar. As expected, the oracle method has the smallest RMSE and OLS the

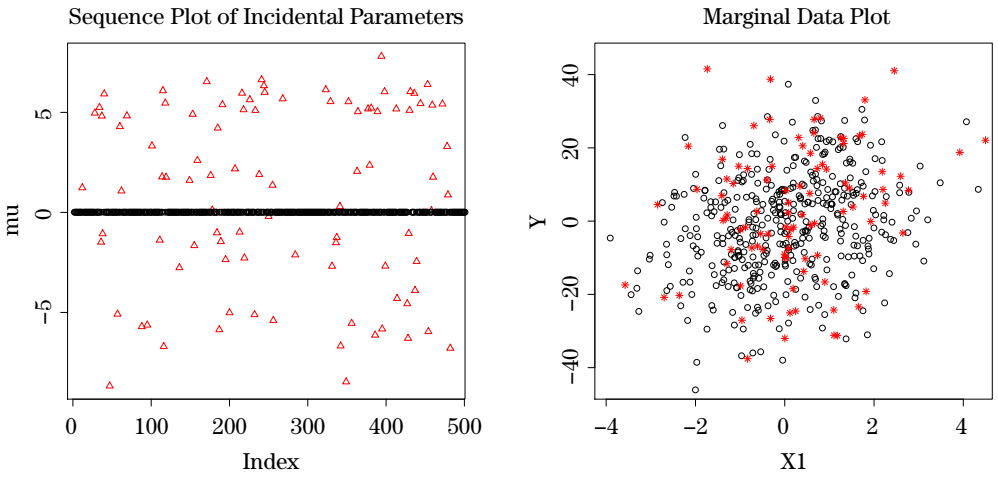

Figure 1. The sequence plot on the left shows 500 incidental parameters $\mu_i^\star$'s with $c = 3$ and $p_w = 0.75$. The (red) triangles are non-zero $\mu_i^\star$'s. The scatter plot on the right shows the responses Y_i's against the first covariate X_{i1}'s of a data set generated with those 500 incidental parameters. The (red) asterisks stand for the contaminated sample points, the ones with nonzero $\mu_i^\star$'s.

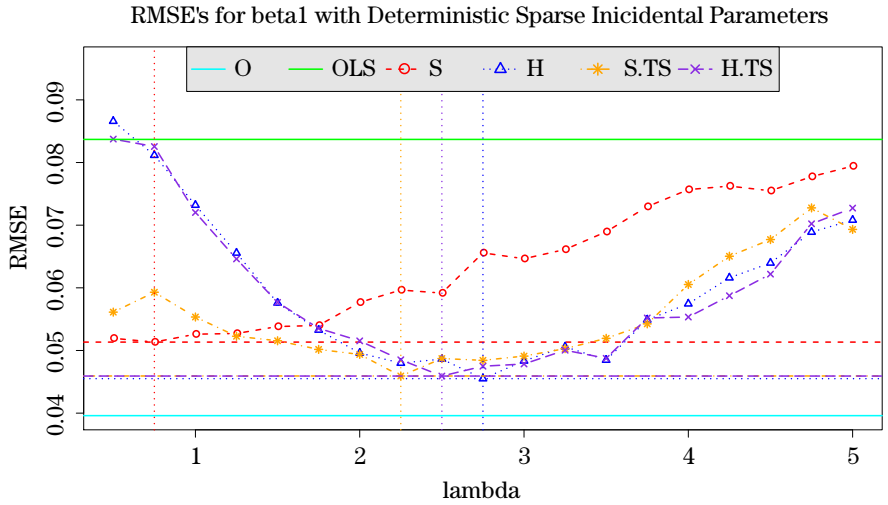

Figure 2. The RMSE's of the O (Oracle), OLS, S (PLS.Soft), H (PLS.Hard), S.TH (PLS.Soft.TwoStep) and H.TH (PLS.Hard.TwoStep) estimators of $\beta_1^\star$ with the incidental parameters shown in Figure 1. The top and bottom solid horizontal lines show the RMSE's for OLS and O, respectively. Other four horizontal lines indicate the minimal RMSE's for those four PLS methods and the corresponding four best λ's are shown by the vertical dotted lines.

largest. The RMSE of each PLS method as a function of λ forms a convex curve, which achieves a minimal RMSE significantly below the line of OLS and close to the line of O. More specifically, the RMSE of PLS.Hard achieves the minimal RMSE when λ is around 2.75. The RMSE of PLS.Soft decreases a little until λ is around 0.75, then increases and stays above the RMSE of PLS.Hard. This reflects the fact that a large λ in a soft-threshold method usually causes a bias. PLS.Hard.TwoStep has a similar performance as PLS.Hard for all λ. PLS.Soft.TwoStep has a similar performance as PLS.Soft when λ is small. As λ becomes large, PLS.Soft.TwoStep moves closer to PLS.Hard than to PLS.Soft, because PLS.Soft.TwoStep and PLS.Hard.TwoStep have similar estimation when λ is large. The minimal RMSE of PLS.Soft is slightly larger than those of other PLS Methods.

Table 1 depicts the minimal RMSE's of the estimators for $\beta_1^\star$ and $\beta_2^\star$ with the corresponding optimal λ's and biases. The biases are ignorable compared with the RMSE's. The optimal λ's for PLS.Soft and other PLS methods are around 0.75 and 2.5, respectively. This indicates the simple soft threshold method tends to work best with a small λ due to the bias issue. Denote the empirical relative efficiency (ERE) of an estimator A with respect to another estimator B as RMSE(B)/RMSE(A). Then, for the estimation of $\beta_1^\star$, the ERE's of PLS.Soft, PLS.Hard, PLS.Soft.TwoStep and PLS.Hard.TwoStep with respect to O are around 78%, 87%, 87%, and 87%, respectively; the ERE's of the PLS methods with respect to OLS are around 176%, 196%, 196%, and 196%, respectively. The ERE's for $\beta_2^\star$ are similar. Thus, in terms of ERE (and RMSE), the PLS methods perform roughly as O and significantly better than OLS.

From Table 1, we also see that the RMSE's of the estimators of $\beta_1^\star$ are smaller than those of $\beta_2^\star$, because the first covariate is less correlated with other covariates than the second one.

To examine the performance of the methods with general incidental parameters, not just those in Figure 1, we generated $\boldsymbol{\mu}^\star$ randomly for each iteration. The iteration number for each simulation was also 1,000.

Figure 3 shows the RMSE's of six estimators of $\beta_1^\star$ with $p_w = 0.5$ and $c = 1$ or 5. Each plot in Figure 3 presents a similar pattern as in Figure 2. When p_w is fixed at 0.5, the RMSE's of each non-oracle estimator of $\beta_1^\star$ increases as the contamination parameter c increases from 1 to 5. This indicates that each non-oracle estimator performs worse as the data becomes more contaminated. However, the PLS estimators are more robust than OLS, which is very sensitive to the change of c. We have also done simulations with $p_w = 0.75$ and the RMSE's of the

Table 1. RMSE's of the Oracle, OLS and LAD estimators and The minimal RMSE's of the penalized estimators of $\beta_1^\star$ and $\beta_2^\star$ with the corresponding optimal or data-driven λ's and biases when the incidental parameters shown in Figure 1 are used. For a data-driven method, the data-driven λ is different in each iteration so that the reported λ's are averages. The lines over the numbers emphasize the numbers are averages. The standard deviations for the data-driven λ's of S.P and H.P are 0.37 and 0.34, respectively.

	O	OLS	S	H	S.TS	H.TS	S.P	H.P	LAD
Bias($\hat{\beta}_1$)$\times 10^4$	0.8	17.8	1.3	0.6	5.9	20.4	9.5	11.5	10.4
RMSE($\hat{\beta}_1$)$\times 10^2$	4.0	8.4	**5.1**	**4.6**	**4.6**	**4.6**	***5.4***	***5.0***	***5.4***
λ			0.75	2.75	2.25	2.5	$\overline{2.45}$	$\overline{2.47}$	
Bias($\hat{\beta}_2$)$\times 10^4$	-2.3	-46.7	24.6	43.5	-21.8	10.0	32.3	13.9	-7.3
RMSE($\hat{\beta}_2$)$\times 10^2$	4.4	9.0	**5.3**	**4.9**	**5.0**	**4.9**	***5.8***	***5.5***	***5.9***
λ			0.75	2.75	2.75	2.5	$\overline{2.45}$	$\overline{2.47}$	

estimators of $\beta_1^\star$ are similar to those with $p_w = 0.75$, so that the corresponding plots are similar to those in Figure 3. The RMSE's of all estimators are stable with respect to p_w, so the magnitudes of the nonzero incidental parameters matter most, not their signs. Some penalized methods perform closely to or even outperform the oracle one when c is small as shown in the plot with $c = 1$, because O ignores all the contaminated data points, even those with very light contamination, while the penalized methods exploit information in such points.

Table 2 contains the RMSE's of the estimators of $\beta_1^\star$ with $p_w = 0.5$ or 0.75 and $c = 0.5, 1, 3$ or 5. For each p_w, as c increases from 0.5 to 5, the RMSE's of O times 10^2 is constantly around 4, those of OLS increase from about 4 to 8.5, and those of PLS grow from about 4 to 5; this affirms the robustness of the PLS estimators. When $c \leq 1$ is small with respect to the variance of random error $\sigma = 1$, the data points are only slightly contaminated, and OLS and PLS methods perform similarly to O. However, when $c \geq 3$ is large, the data are more contaminated, and the RMSE's of OLS are significantly larger, but PLS methods perform closely to O.

4.2. Performance of data-driven penalized methods

Previous simulations have shown that the PLS methods with optimal λ's have good RMSE's compared with those of the Oracle and OLS. In practice, optimal λ's are unknown. An approach to obtaining a data-driven λ has been introduced in Subsection 3.2. Since, as shown in previous simulation results, the two-step PLS methods perform similarly as the one-step PLS methods, PLS.Soft and PLS.Hard, only the latter were studied by simulations with data-driven

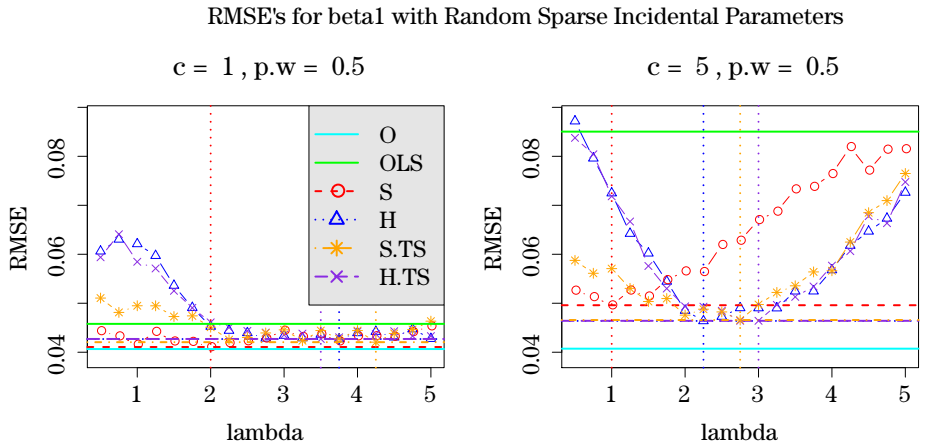

Figure 3. Similar to Figure 2, these plots show the RMSE's of the O (Oracle), OLS, S (PLS.Soft), H (PLS.Hard), S.TH (PLS.Soft.TwoStep) and H.TH (PLS.Hard.TwoStep) estimators of $\beta_1^\star$ with randomly generated $\boldsymbol{\mu}^\star$ under two settings with $p_w = 0.5$ and $c = 1$ or 5.

Table 2. Similar to Table 1, this one shows the RMSE's and minimal RMSE's of nine estimators of $\beta_1^\star$ under eight settings on randomly generated $\boldsymbol{\mu}^\star$ with $p_w = 0.5, 0.75$ and $c = 0.5, 1, 3, 5$. The standard deviations of the data-driven λ's of S.P and H.P for different settings are between 0.2 and 0.45.

RMSE($\hat{\beta}_1$)$\times 10^2$	O	OLS	S	H	S.TS	H.TS	S.P	H.P	LAD
$(p_w, c) = (0.5, 0.5)$	4.06	4.06	3.92	3.95	3.90	3.96	4.01	4.38	4.85
λ			1.25	3.5	2.75	2.5	2.08	2.15	
$(p_w, c) = (0.5, 1)$	4.06	4.58	4.11	4.27	4.21	4.27	4.01	4.59	4.75
λ			2	3.75	4.25	3.5	2.20	2.24	
$(p_w, c) = (0.5, 3)$	4.12	6.47	**4.81**	**4.99**	**4.81**	**4.80**	***5.64***	***5.19***	***5.49***
λ			1	2.5	2	2.25	2.42	2.45	
$(p_w, c) = (0.5, 5)$	4.07	8.50	**4.96**	**4.63**	**4.66**	**4.64**	***6.36***	***4.87***	***5.52***
λ			1	2.25	2.75	3	2.52	2.58	
$(p_w, c) = (0.75, 0.5)$	4.13	4.02	3.91	3.88	3.98	3.91	4.08	4.41	4.79
λ			1.5	5	3.25	4.5	2.08	2.14	
$(p_w, c) = (0.75, 1)$	4.17	4.41	4.19	4.15	4.15	4.20	4.16	4.59	4.98
λ			3.25	3	4	3.5	2.15	2.23	
$(p_w, c) = (0.75, 3)$	4.15	5.99	**4.91**	**4.93**	**4.80**	**5.02**	***5.35***	***5.06***	***5.52***
λ			1	2.25	2	2.5	2.44	2.47	
$(p_w, c) = (0.75, 5)$	3.97	8.41	**5.01**	**4.66**	**4.75**	**4.66**	***6.22***	***4.90***	***5.76***
λ			1.5	2.25	2.5	3	2.55	2.60	

λ's, and they are denoted by PLS.Soft.Prac (S.P) and PLS.Hard.Prac (H.P), respectively. For estimating data-driven λ's, let $\alpha_l = 2$ and $\alpha_u = 7$. The size of

the pure data set n_{pure} was $n/2$ and that of the testing data set was $n_{pure}/2$.

Simulations were first run with the deterministic sparse incidental parameters as shown in Figure 1. In Table 1, the RMSE's of estimators of $\beta_1^\star$ from PLS.Soft.Prac and PLS.Hard.Prac are around 5.4 and 5.0, slightly larger than the optimal values 5.1 and 4.6, respectively. However, they are still significantly smaller than the RMSE of OLS 8.4. The observations of the estimators of $\beta_2^\star$ are similar. We also evaluated the performance of the data-driven PLS methods with random sparse incidental parameters. Table 2 shows that, for a given p_w, when c is 0.5 and 1 (small), the RMSE's of PLS.Soft.Prac and PLS.Hard.Prac are close to those of PLS.Soft and PLS.Hard with the optimal λ's. In these cases, PLS.Soft.Prac performs slightly better than PLS.Hard.Prac, and even better than PLS.Soft with the optimal λ and the Oracle method. On the other hand, for a given p_w, when c is 3 and 5 (large), the RMSE's of PLS.Soft.Prac and PLS.Hard.Prac are greater than those of PLS.Soft and PLS.Hard, respectively, but still less than those of OLS. In these cases, the RMSE's of PLS.Soft.Prac are larger than those of PLS.Hard.Prac, which indicates the bias issue with the soft threshold method. Thus, the data-driven regularization parameter works well with penalized estimation. When the data is slightly contaminated, the soft penalty is preferred; otherwise, the hard penalty is recommended.

Tables 1 and 2 also contain the RMSE of the least absolute deviation regression method (LAD) used in Fan, Feng and Tong (2012) with all but not part of the sample points with small residuals. Generally speaking, in both deterministic and random incidental parameter cases, LAD performs similarly to the PLS methods with data-driven λ's. More specifically, when $c \leq 1$ is small, PLS.Soft.Prac outperforms LAD; otherwise, LAD performs better. For all cases, LAD is dominated by PLS.Hard.Prac. These observations affirm that LAD is an effective robustness method and the penalized methods make improvement.

4.3. Data-driven confidence intervals

We next turn to the finite-sample performance of the asymptotic confidence interval (CI) (3.2) for $\beta_j^\star$ with $j = 1$ and 2, based on PLS two-step methods. Since (3.2) is based on the properties of the penalized two-step estimator with the soft penalty, we focused on PLS.TS.Soft with a data-driven regularization parameter λ. The choice of λ in Subsection 3.2 for minimizing RMSE is usually no longer suitable for constructing confidence intervals, since it is designed to achieve minimal RMSE. We first obtained $\hat{\sigma}_{pure}$ as in the data-driven procedure in Subsection 3.2 and then simply set the data-driven λ to be five times $\hat{\sigma}_{pure}$.

Table 3. Coverage rates (CR) and average length (AL) of 95% confidence intervals for $\beta_1^\star$ and $\beta_2^\star$ from O, OLS, PLS.TwoStage.Soft.Prac methods under three settings on deterministic sparse incidental parameters.

(p_1, p_2)	$\beta_1^\star$	O	OLS	S.TS.P	$\beta_2^\star$	O	OLS	S.TS.P
(0.01, 0.01)			0.950	0.944			0.948	0.945
(0.03, 0.03)	CR	0.955	0.951	0.948	CR	0.948	0.946	0.947
(0.05, 0.05)			0.946	0.945			0.949	0.946
(0.01, 0.01)			0.200	**0.135**			0.213	**0.144**
(0.03, 0.03)	AL	0.133	0.279	**0.137**	AL	0.142	0.297	**0.146**
(0.05, 0.05)			0.382	**0.139**			0.407	**0.149**

Since $\hat{\sigma}_{pure}$ tends to underestimate σ, this data-driven λ is usually not large with respect to σ. Denote this method as PLS.TwoStage.Soft.Prac or S.TS.P. After plugging in $\hat{\sigma}$ and $\hat{\sigma}_j^{-1}$, the square root of the (j, j) element of $\hat{\mathbf{\Sigma}}_X^{-1}$, and replacing n by $m = \#(\hat{I}_0)$ at (3.2), we obtained the data-driven CI $[\tilde{\beta}_j \pm m^{-1/2}\hat{\sigma}\hat{\sigma}_j^{-1}z_{\alpha/2}]$, where $\tilde{\beta}_j$ is the PLS.TwoStage.Soft.Prac estimator of $\beta_j^\star$ for each j.

This data-driven CI was compared with CI's based on Oracle and OLS methods. More specifically, write the Oracle and OLS estimators of $\beta_j^\star$ as $\hat{\beta}_j^{(O)}$ and $\hat{\beta}_j^{(OLS)}$, respectively. Then, the corresponding CI's are given by $[\hat{\beta}_j^{(O)} \pm m_o^{-1/2}\hat{\sigma}^{(O)}\hat{\sigma}_j^{-1}z_{\alpha/2}]$ and $[\hat{\beta}_j^{(OLS)} \pm n^{-1/2}\hat{\sigma}^{(OLS)}\hat{\sigma}_j^{-1}z_{\alpha/2}]$, where m_o is the number of zero incidental parameters and $\hat{\sigma}^{(O)}$ and $\hat{\sigma}^{(OLS)}$ are the estimators of σ from O and OLS methods, respectively.

The simulation settings were the same as the previous ones with deterministic sparse incidental parameters, except the following changes. (a) The number of covariates d was reduced to 5 from 50, because with $d = 50$ and level 95%, even the empirical coverage rate (CR) of the oracle confidence interval for $\beta_1^\star$ is 93.5%, not very close to 95%. (b) The iteration number was increased from 1,000 to 10,000 to improve the accuracy of CR's. (c) The probabilities of nonzero incidental parameters (p_1, p_2) were set to be $(0.01, 0.01)$, $(0.03, 0.03)$ and $(0.05, 0.05)$; the contamination parameter c was increased to 10. In order to achieve good second-order asymptotic approximation, one could either increase the sample size or enlarge the signal noise ratio. We adopted the latter.

Table 3 reports the empirical coverage rates (CR) and average lengths (AL) of the CI's of $\beta_1^\star$ and $\beta_2^\star$ from O, OLS and PLS.TS.Soft.Prac methods under three different settings on the incidental parameters. For the oracle method, these three settings are the same, and thus only one set of simulation results is presented. Table 3 shows that the CR's of all methods under all settings

are close to the nominal level 0.95. The OLS treats the deterministic incidental parameters as random and achieves excellent CR's. However, the AL's of OLS are significantly larger than those of O and PLS.TS.Soft.Prac, especially when there are more non-zero incidental parameters. On the other hand, the AL's of PLS.TS.Soft.Prac are only slightly larger than those of O. This means PLS.TS.Soft.Prac has excellent efficiency in terms of AL's given excellent CR's. The AL's for $\beta_1^\star$ are less than those for $\beta_2^\star$, because the asymptotic variance of $\hat{\beta}_1$ is less than that of $\hat{\beta}_2$ when the covariance matrix $\mathbf{\Sigma}_X$ is a Toeplitz matrix. Simulations with random incidental parameters under the same settings have also been done and the results are similar to those in Table 3, with slightly inflated AL's for OLS and PLS.TS.Soft.Prac due to the randomness of the incidental parameters.

4.4. Data analysis

We implemented penalized estimation with the soft penalty in the method of estimating the false discovery proportion of a multiple testing procedure of Fan, Feng and Tong (2012) for investigating the association between the expression level of gene CCT8, closely related to Down Syndrome phenotypes, and thousands of SNPs. The data set consists of three populations: 60 Utah residents (CEU), 45 Japanese and 45 Chinese (JPTCHB), and 60 Yoruba (YRI). More details on the data set can be found in Fan, Feng and Tong (2012).

In their method, a filtered least absolute deviation regression (LAD) is used to estimate the loading factors with 90% of the cases (SNPs) whose test statistics are small and thus the resulting estimator is statistically biased. We upgraded this step with S.P as described in Subsections 3.2 and 4.2, and re-estimated the number of false discoveries $V(t)$, and the false discovery proportion $\mathrm{FDP}(t)$ as functions of $-\log_{10}(t)$, where t is a thresholding value. Figure 4 shows the number of total discoveries $R(t)$, $\hat{V}(t)$ and $\widehat{\mathrm{FDP}}(t)$ from procedures using filtered LAD and S.P. It is clear that $\hat{V}(t)$ and $\widehat{\mathrm{FDP}}(t)$ with S.P are uniformly larger than but reasonably close to those with filtered LAD. Table 4 contains $R(t)$ and $\widehat{\mathrm{FDP}}(t)$ with filtered LAD and S.P for several specific thresholds. The estimated FDPs with S.P for CEU and YRI are slightly larger than those with LAD and $\widehat{\mathrm{FDP}}$ for JPTCHB with S.P is more than twice that with filtered LAD. This suggests that the estimation of FDP with filtered LAD tends to be optimistic.

5. Conclusion and Discussion

This paper considers the estimation of structural parameters with the pres-

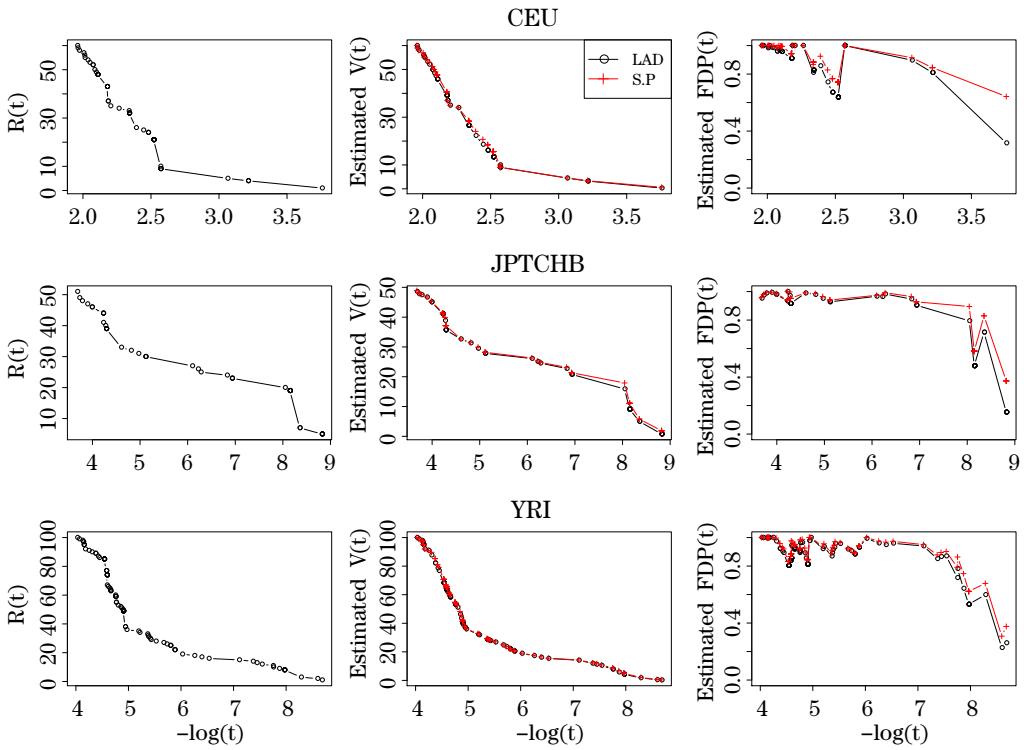

Figure 4. Discovery number $R(t)$, estimated false discovery number $V(t)$ and estimated false discovery proportion $\mathrm{FDP}(t)$ as functions of a threshold on t for populations CEU, JTPCHB and YRI. The x-axis is $-\log_{10}(t)$.

Table 4. Discovery numbers $R(t)$ and estimated false discover proportions $\widehat{\mathrm{FDP}}(t)$s from methods with LAD and S.P for specific values of threshold t.

Population	t	$R(t)$	$\widehat{\mathrm{FDP}}(t)$ with LAD	$\widehat{\mathrm{FDP}}(t)$ with S.P
CEU	6.12×10^{-4}	4	0.810	0.845
JPTCHB	1.51×10^{-9}	5	0.153	0.373
YRI	2.54×10^{-9}	2	0.227	0.308

ence of sparse incidental parameters in a linear regression model. By exploiting the sparsity, we propose an estimation method penalizing the incidental parameters. The penalized estimator of the structural parameters is consistent and asymptotically Gaussian and achieves an oracle property. On the contrary, the penalized estimator of the incidental parameters possesses only partial selection consistency but not consistency. Thus, the structural parameters are consistently estimated while the incidental parameters are not, which presents a partial con-

sistency phenomenon. To construct better confidence regions for the structural parameters, we propose a two-step estimator that has fewer possible asymptotic distributions and can be asymptotically even more efficient than the one-step penalized estimator when the size and magnitude of nonzero incidental parameters are substantially large.

Simulations show that the penalized methods with best regularization parameters achieve significantly smaller mean square errors than the ordinary least squares method that ignores the incidental parameters. We provide a data-driven regularization parameter, with which the penalized estimators continue to significantly outperform ordinary least squares when the nonzero incidental parameters are too large to be neglected. In terms of average length and coverage rate, the advantage of the confidence intervals based on the two-step estimator with a data-driven regularization parameter is verified by simulations. A data set on genome-wide association study is analyzed with a multiple testing procedure and false discovery proportions are estimated with the help of the data-driven penalized method.

In econometrics, a fixed effect panel data model is given by, for $1 \le i \le n$ and $1 \le t \le T$,

$$Y_{it} = \mu_i^\star + \boldsymbol{X}_{it}^T \boldsymbol{\beta}^\star + \epsilon_{it}, \tag{5.1}$$

where $\mu_i^\star$'s are unknown fixed effects. When T diverges, the fixed effects can be consistently estimated. When T is finite and greater than or equal to 2, although the fixed effects can no longer be consistently estimated, they can be removed by a within-group transformation: for each i, $Y_{it} - \bar{Y}_i = (\boldsymbol{X}_{it} - \bar{\boldsymbol{X}}_i)^T \boldsymbol{\beta}^\star + \epsilon_{it} - \bar{\epsilon}_i$, where $\bar{Y}_i$, $\bar{\boldsymbol{X}}_i$ and $\bar{\epsilon}_i$ are the averages of Y_{it}'s, $\boldsymbol{X}_{it}$'s and ϵ_{it}'s, respectively. When T is 1, however, the within-group transformation fails. For this case, model (5.1) becomes model (1.1) so that the proposed penalized estimations provide a solution under the sparsity assumption on the fixed effects.

Although this paper only illustrates the partial consistency phenomenon of a penalized estimation method for a linear regression model, such a phenomenon shall universally exist for a general parametric model that contains both structural parameters and high-dimensional sparse incidental parameters. For example, consider a panel data logistic regression model: $P(Y_{it} = 1|\boldsymbol{X}_{it}) = \{1 + \exp(-(\mu_i^\star + \boldsymbol{X}_{it}^T \boldsymbol{\beta}^\star))\}^{-1}$ with $1 \le t \le T$. When T is finite, the fixed effects $\mu_i^\star$'s cannot be removed by the within-group transformation as in the panel data linear model (5.1). However, the proposed penalized estimations can still provide a solution.

Finally, if the number of the structural parameters diverging faster than the sample size and they are also sparse, it is expected that the partial consistency phenomenon will continue to appear when the sparsity penalty is imposed on both the structural parameters and the incidental ones.

Supplementary Materials

The proofs of the theoretical results and additional materials for Sections 1 to 3 are available in a online supplementary file, which also contains an extension case with the number of covariates growing with, but slower than, the sample size.

Acknowledgment

The authors thank the Editor, an associate editor, and three referees for their helpful comments that have resulted in significant improvements of this paper. The research was partially supported by NSF grants DMS-1206464 and DMS-0704337 and NIH Grants R01-GM100474-01 and R01-GM072611-08.

References

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* **72**, 355–366.

Bean, D., Bickel, P., El Karoui, N., Lim, C. and Yu, B. (2012). Penalized robust regression in high-dimension. Technical Report, Department of Statistics, University of California, Berkeley.

Chen, X. H., Wang, Z. J. and McKeown, M. J. (2010). Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transactions on Information Theory* **56**, 5131–5149.

Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics* **42**, 324–351.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society Series B. Statistical Methodology* **74**, 745–771.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39**, 3320–3356.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics* **1**, 799–821.

Jahn, J. (2007). *Introduction to the Theory of Nonlinear Optimization*. Springer Berlin Heidelberg.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* **27**, 887–906.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics* **5**, 1015–1053.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* **95**, 391–413.

Moreira, M. J. (2009). A maximum likelihood method for the incidental parameter problem. *The Annals of Statistics* **37**, 3660–3696.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Portnoy, S. and He, X. (2000). A robust journey in the new millennium. *Journal of the American Statistical Association* **95**, 1331–1335.

Shiryaev, A. N. (1995). *Probability*. 2nd Edition. Springer-Verlag.

Tang, R., Banerjee, M. and Kosorok, M. R. (2012). Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *The Annals of Statistics* **40**, 45–72.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B. Statistical Methodology* **58**, 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7**, 2541–2563.

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: jqfan@princeton.edu

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA.

E-mail: tang.runlong@gmail.com

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: shi.xiaofeng@alumni.princeton.edu

(Received January 2017; accepted October 2007)