**BIOMETRIC METHODOLOGY**

# Drawing inferences for High-dimensional Linear Models: A Selection-assisted Partial Regression and Smoothing Approach†

**Zhe Fei[1], Ji Zhu[2], Moulinath Banerjee[2], and Yi Li[1,*]**

[1] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

[2] Department of Statistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

**email*: yili@umich.edu

This paper has been submitted for consideration for publication in *Biometrics*



**Additional Supporting Information may be found in the online version of this article.**

**Received 31 March 2018; Revised 19 November 2018; Accepted 6 December 2018**
**Biometrics**

**DOI 10.1111/biom.13013**

 

Summary: Drawing inferences for high-dimensional models is challenging as regular asymptotic theories are not applicable. This paper proposes a new framework of simultaneous estimation and inferences for high-dimensional linear models. By smoothing over partial regression estimates based on a given variable selection scheme, we reduce the problem to a low-dimensional least squares estimation. The procedure, termed as Selection-assisted Partial Regression and Smoothing (SPARES), utilizes data splitting along with variable selection and partial regression. We show that the SPARES estimator is asymptotically unbiased and normal, and derive its variance via a nonparametric delta method. The utility of the procedure is evaluated under various simulation scenarios and via comparisons with the de-biased LASSO estimators, a major competitor. We apply the method to analyze two genomic datasets and obtain biologically meaningful results. 

Key words: Confidence intervals; High-dimensional inference; Hypothesis testing; Multisample-splitting; Selection-assisted Partial Regression and Smoothing (SPARES).

## 1. Introduction

Consider the classical linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta^0 + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{Y} = (y_1, y_2, ..., y_n)^{\mathrm{T}}$ is the $n$-vector of the response variable; $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ is the $n \times p$ design matrix that consists of $p$ covariate vectors $X_j$'s; $\boldsymbol{X}$ can also be written as $\boldsymbol{X} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, .., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ represents the $p$-vector of covariates for the $i^{th}$ individual; $\beta^0 = (\beta_1^0, ..., \beta_p^0)^{\mathrm{T}}$ is the true parameter vector of interest; $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^{\mathrm{T}}$ is the random noise vector and $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n$.

In the traditional low-dimensional setting when $n > p$, it is well known that least squares estimator $\hat{\beta}_{LS} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ converges to a normal distribution centered at $\beta^0$, which provides exact estimation and inferences through explicitly computable p-values and confidence intervals. On the other hand, when $n < p$, the least squares estimation would fail because the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ is singular. However the $n < p$ problem has become increasingly relevant over the past two decades with the common availability of high-throughput data. The goal is often to find a parsimonious model to explain the response in the presence of massive covariates. A number of selection and estimation methods including LASSO (Tibshirani, 1996), Adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001), ISIS (Fan and Lv, 2008), among others, are available.

More recently, interest in the statistical community has shifted to making reliable inferences in high-dimensional models. Researchers have been trying to tackle the problem from different angles. One direction is to make inferences based on the *selected* model, i.e. the one that is chosen by a given variable selection procedure. Wasserman and Roeder (2009) proposes a multi-stage procedure that is based on data splitting to separate selection and inference; Berk et al. (2013) provides conservative confidence intervals for the selected variables by defining a set of candidate models; Lee and Taylor (2014); Lee et al. (2016) develops the conditional

asymptotics of the coefficient estimates, given the selected model. The second direction is to estimate and make inferences of the low-dimensional parameters in the high dimensional models. Belloni et al. (2013, 2014) propose a double selection procedure instead of a single selection step to estimate and construct confidence regions for a regression parameter of primary interest. Some other works propose estimators and inferences based on penalized estimation. A typical example is the bias correction method based on LASSO (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014), which provides point estimation and confidence intervals for the model parameters. There is also work by Ning and Liu (2017) that proposes hypothesis tests and confidence regions based on the decorrelated score function and test statistic.

These approaches have their merits and demerits. While Wasserman and Roeder (2009); Lee and Taylor (2014); Lee et al. (2016) aim at exact inference for post-selection estimates, it is confined to the selected model from the *"first step."* Thus, flaws in the initial model-selection step, cannot be rectified in subsequent steps. The limitation of requiring perfect model selection is improved in Belloni et al. (2014), meanwhile, Wasserman and Roeder (2009); Meinshausen et al. (2009) recommend *not* performing selection and estimation on the same data set. On the other hand, the performance of the original de-biased LASSO estimator relies heavily on the accuracy of estimating the precision matrix, i.e. $\mathbf{\Sigma}^{-1}$, which plays an unduly crucial role in the estimation and inference subsequently. In Javanmard and Montanari (2014), they relaxed the required accuracy of estimating $\mathbf{\Sigma}^{-1}$ (the matrix $M$ in their paper), instead they set $M$ as to minimize the error term and the variance of the target Gaussian limit.

In this paper we propose a novel approach to consistently estimate $\beta^0$, provide p-values for all covariates, and compute confidence intervals for any fixed subset of parameters in high-dimensional linear models. The approach, coined *Selection-assisted Partial Regression*

*and Smoothing* (SPARES), possesses asymptotic unbiasedness and asymptotic normality. Our idea takes advantage of the multisample-splitting method in Meinshausen et al. (2009), which defines a p-value for each predictor from each sample-splitting and then aggregates these p-values to declare a single p-value per feature. One possible criticism of this approach is that the p-values and the aggregation have a certain arbitrary angle to them: for example, features not selected in each sample-split subsample are all assigned a p-value 1. In contrast, our SPARES estimator utilizes partial regression to estimate $\beta^0$ in each sample-split followed by a natural smoothing step. In each data split, our procedure provides an estimate of $\beta_j^0,\ j = 1, 2, .., p$ regardless of whether it was chosen by the selection procedure. Such idea of attaching variable $j$ to the selected variables is also used in Belloni et al. (2014). Then we average over the variation of the selection and sample-split to obtain a smoothed estimator. For these reasons, SPARES is *not* a post model-selection method. Furthermore, our approach avoids the need to estimate the high-dimensional precision matrix.

Our approach stands out from the majority of related works (Wasserman and Roeder, 2009; Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014; Belloni et al., 2014; Ning and Liu, 2017) in that it is neither restricted to a fixed realization of the selected model nor limited to a certain selection procedure. The smoothing accomplished through multisample-splitting ensures that the $\widehat{\beta}_j$'s are asymptotically normal with negligible bias while the standard errors can be readily estimated via a nonparametric delta method (Efron, 2014). Consequently, inferences can be made for each and every $\beta_j^0, j = 1, 2, .., p$ without having to confront the curse of dimensionality. As shown in the data applications, our method is advantageous in giving uncertainty measures (such as p-values) to all high dimensional coefficients at once.

The rest of this paper is organized as follows. Section 2 describes the SPARES estimator and Section 3 develops its theoretical properties. Section 4 shows how to draw inferences

through SPARES, including confidence intervals and significance tests. Section 5 discusses the extension to a subvector of $\beta^0$ with a fixed dimension. In Section 6 we conduct simulations to examine the performance of SPARES and present comparisons to de-biased LASSO methods. Section 7 comprises two real data applications and Section 8 summarizes the merit of this work and pinpoints future research.

## 2. Proposed Method

Let $[p] = \{1, 2, .., p\}$ denote the set of integers for any positive $p$. For a vector $V$ of length $p$, denote the entry corresponding to subscript $j \in [p]$ by $V_j$ or $(V)_j$; for a square matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_{p\times p}$, denote the entry corresponding to subscripts $j, k \in [p]$ by $\mathbf{\Sigma}_{jk}$ or $(\mathbf{\Sigma})_{jk}$ for clarity if necessary; for a subset $S \subset [p]$, denote the sub-design matrix $X_S = (X_j)_{j\in S}$ and the sub-covariance matrix $\mathbf{\Sigma}_S = (\mathbf{\Sigma}_{jk})_{j,k\in S}$. The projection matrix of $X_S$ is denoted as $H_S = X_S(X_S^{\mathrm{T}}X_S)^{-1}X_S^{\mathrm{T}}$. The active set of $\beta^0$ is $S_{0,n} = \{j \in [p] : \beta_j^0 \neq 0\}$.

**One-time SPARE:** We first introduce the estimation of $\beta^0$ through Selection-assisted Partial Regression (SPARE) on a single data-split. Given data $D_n = (\boldsymbol{X}, \boldsymbol{Y})$ as in model (1) and a generic selection procedure $\mathcal{S}_\lambda$ with parameter $\lambda$, we first split $D_n$ into two halves $D_1$ and $D_2$, with $|D_1| = \lfloor n/2 \rfloor$, $|D_2| = \lceil n/2 \rceil$, the floor and ceiling of it. Denote the subset of variables selected by $\mathcal{S}_\lambda$ on $D_2$ as $S = \mathcal{S}_\lambda(D_2)$. Next on $D_1 = (\boldsymbol{X}^1, \boldsymbol{Y}^1)$, the partial regression estimator for $\beta_j^0$, $j \in [p]$ is

$$\widetilde{\beta}_j = \left\{ ({\boldsymbol{X}^1_{S\cup j}}^{\mathrm{T}} \boldsymbol{X}^1_{S\cup j})^{-1} {\boldsymbol{X}^1_{S\cup j}}^{\mathrm{T}} \boldsymbol{Y}^1 \right\}_j , \tag{2}$$

which is the coefficient estimate corresponding to $\boldsymbol{X}_j^1$ from the least squares regression of $\boldsymbol{Y}^1$ on $\boldsymbol{X}^1_{S\cup j}$. Moreover, (2) can be written as $\widetilde{\beta}_j = \left\{ {\boldsymbol{X}_j^1}^{\mathrm{T}} (I_{n/2} - H^1_{S\setminus j}) \boldsymbol{X}_j^1 \right\}^{-1} {\boldsymbol{X}_j^1}^{\mathrm{T}} (I_{n/2} - H^1_{S\setminus j}) \boldsymbol{Y}^1$ in the partial regression formulation.

Let $S_C = [p] \setminus S$, we can write the one-time SPARE estimator compactly as

$$\widetilde{\beta}(D_1, S) = \begin{pmatrix} \widetilde{\beta}_S \\ \widetilde{\beta}_{S_C} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}_S^{1\,\mathrm{T}} \boldsymbol{X}_S^1)^{-1} \boldsymbol{X}_S^{1\,\mathrm{T}} \boldsymbol{Y}^1 \\ \left[\mathrm{diag}\{\boldsymbol{X}_{S_C}^{1\,\mathrm{T}} (I_{n/2} - H_S^1) \boldsymbol{X}_{S_C}^1\}\right]^{-1} \boldsymbol{X}_{S_C}^{1\,\mathrm{T}} (I_{n/2} - H_S^1) \boldsymbol{Y}^1 \end{pmatrix}. \quad (3)$$

The rationale for SPARE to work is that given a subset of important predictors $S \subset [p]$ that is close to the active set $S_{0,n}$, the partial regression estimator (2) would be a fine estimator that is close to the truth $\beta_j^0$, for all $j \in [p]$. In fact, as long as $S \supset S_{0,n}$, (2) would be an unbiased estimator for $\beta_j^0$, regardless of $j \in S$ or not. However, given the large number of predictors, the one-time SPARE estimator is highly variable, and heavily depends on the selected $S$ and the specific split of data.

**SPARES:** To overcome this difficulty, we introduce its smoothed version, the SPARES estimator, which is derived from multisample-splitting and repeated applications of SPARE. For a large enough $B$ and each $b = 1, 2, .., B$, we first draw a sample of size $n/2$, with replacement, from the full data and denote it as $D_1^b$. When $n$ is odd, we interpret $n/2$ as $\lfloor n/2 \rfloor$. Let $I_1 = \{i_1, i_2, ..., i_{n/2}\}, 1 \leqslant i_k \leqslant n$ be the collection of indices of the observations in $D_1^b$. Next, we collect the observations that are not drawn in $D_1^b$ as $D_2^b$ with index set $I_2 = [n] \setminus I_1$. Thus $I_1 \cup I_2 = [n]$ and $I_1 \cap I_2 = \emptyset$. Now the application of SPARE by (3) is $\widehat{\beta}^b = \widetilde{\beta}(D_1^b, S^b)$, where $S^b = \mathcal{S}_\lambda(D_2^b)$; the final step is to average over all $\widehat{\beta}^b$'s,

$$\widehat{\beta}_{\mathrm{SPARES}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}^b. \quad (4)$$

In terms of the computational cost, each of the one-time SPARE has the same time complexity as one run of LASSO ($O(np^2)$), and the cost of the SPARES procedure is $B$ times that. But with the help of parallel computing, we could largely reduce the computation time by any desired factor $K$ depending on the computing tool. Thus the time complexity of SPARES is $O(Bnp^2/K)$, a multiple of one-time LASSO proportional to the number of re-samples. Empirically the total time cost of the SPARES procedure is linear in $p \log n$.

In the rest of the paper, we will always use $\widetilde{\beta}$ for the one-time SPARE estimator and $\widehat{\beta}$ for

the SPARES estimator. Both the one-time SPARE and the SPARES possess the asymptotic unbiasedness and normality, but SPARES is much more stable due to the smoothing effect from multisample-splitting, which we will explore in depth throughout the rest of this paper.

## 3. Theoretical Results

### 3.1 *One-time SPARE*

We first establish the asymptotic property of the one-time SPARE estimator under the following assumptions.

(A1). Randomness of Data: In model (1), $\varepsilon_i \perp \mathbf{x}_i$; $\varepsilon_i$'s are i.i.d. random errors with mean zero, finite variance $\sigma^2$ and finite third absolute moment $\mathbf{E}|\varepsilon_i|^3 \leqslant \rho_0$; $\boldsymbol{X} = (\mathbf{x}_1^{\mathrm{T}}, ..., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{x}_i$'s are i.i.d. mean zero sub-Gaussian random vectors in $\mathbf{R}^p$ with covariance matrix $\boldsymbol{\Sigma}_{p\times p}$, whose eigenvalues are bounded,

$$0 < c_{\min} \leqslant \lambda_{\min}(\boldsymbol{\Sigma}) \leqslant \lambda_{\max}(\boldsymbol{\Sigma}) \leqslant c_{\max} < \infty.$$

$\mathbf{x}_i$'s also have finite component-wise third absolute moments $\forall j$, $\mathbf{E}|x_{ij}|^3 \leqslant \rho_1$.

(A2). Order of Model Parameters: There exist constants $0 < c_1 \leqslant 1, c_\beta > 0$ such that $s_0 = |S_{0,n}| = O(n^{c_1})$, $\max_j |\beta_j^0| \leqslant c_\beta$.

(A3). Sure Screening Property: There exists a sequence $\{\lambda_n\}_{n\geqslant 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\widehat{S}_{n,\lambda_n}|/n \leqslant \eta$, and

$$P(\widehat{S}_{n,\lambda_n} \supset S_{0,n}) \geqslant 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty. \tag{5}$$

Here $\widehat{S}_{n,\lambda_n}$ denotes the selected set of variables with sample size $n$ and tuning parameter $\lambda_n$.

REMARK 1: The sure screening property is met in Fan and Lv (2008); Fan and Song (2010), and is guaranteed with the right order of tuning parameter $\lambda$ using LASSO (Bach, 2008). More specifically, by Fan and Lv (2008); Fan and Song (2010), in addition to assump-

tions (A1) and (A2), the following conditions are required for the sure screening property to hold:

- $\mathrm{Var}(\boldsymbol{Y}) = O(1)$, and for some $\kappa \geqslant 0$ and $c_0, c_3 > 0$, $\min_{j \in S_0} |\beta_j^0| \geqslant c_0/n^\kappa$ and

$$\min_{\beta_j \neq 0} \left|\mathrm{cov}\left(\beta_j^{-1}\boldsymbol{Y}, \boldsymbol{X}_j\right)\right| \geqslant c_3;$$

- $\log p = O(n^\xi)$ for some $0 < \xi < 1 - 2\kappa$.

When $\kappa \geqslant 1/3$, the sparsity requirement implied by Fan and Lv (2008), $s_0 = o(n^\theta)$ for some $0 < \theta < 1 - 2\kappa$, is stronger than that in Javanmard and Montanari (2018), which is $s_0 = o(n/(\log p)^2)$. When $\kappa < 1/3$, the comparison between the two conditions are inconclusive. Please see conditions 1-4 in Fan and Lv (2008) for more details.

In (A1), only a moment condition is required on the error terms and a sub-Gaussian distribution for the covariates. For comparisons, while the asymptotic normality of the whole $p-$dimensional de-biased estimator is not guaranteed for non-Gaussian errors, a central limit theorem argument can be used to obtain approximate Gaussianity of components of fixed dimension (Bühlmann et al., 2014). Thus the inference for any fixed low-dimensional parameter is still valid for these types of methods under sub-Gaussian errors with finite moment conditions. In (A2), there is no direct assumption on the order of $p$, however, it is implied through (A3), a condition made directly on the selection method. One reason for such an assumption, instead of more basic ones like the order of $p$ or the covariance structure of the predictors, is that selection only plays an assistive role in our method; the estimation part is in fact low-dimensional and therefore does not directly require typical high-dimensional conditions.

THEOREM 1: *Given model (1) and assumptions (A1)-(A3), consider the one-time SPARE estimator* $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, .., \widetilde{\beta}_p)^{\mathrm{T}}$ *as defined in (3). Denote* $m = \lfloor n/2 \rfloor$, $\widetilde{\sigma}_j^2 = \sigma^2 \left( {\boldsymbol{X}^1_{S\cup j}}^T X^1_{S\cup j}/m \right)^{-1}_{jj}$.

*Then* $\forall j \in [p]$, *as* $m \to \infty$,

$$\sqrt{m}(\widetilde{\beta}_j - \beta_j^0)/\widetilde{\sigma}_j \to N(0,1). \tag{6}$$

REMARK 2: Note that we could always let the quantity of interest in (6) to be zero whenever $S_0 \not\subset S$, whose probability goes to zero by (A3). Thus we only need to show the convergence when the event $S_0 \subset \widehat{S}$ holds.

The proof is presented in the Web Appendix A.

### 3.2 *SPARES*

Given the high volume of predictors in the model (1), the one-time estimator is expected to be noisy and unstable, especially for all the $j \notin S_{0,n}$ that are the majority of the $p-$vector $\beta^0$. In contrast, the SPARES estimator is more stable as it smooths over both estimation and selection. As the SPARES introduces extra dependency between the selections $S^b$'s and the partial regression estimates, the following condition, which is stronger than "sure screening", is required for the desired theoretical property.

(B3). Selection Consistency: There exists a sequence $\{\lambda_n\}_{n\geqslant 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\widehat{S}_{n,\lambda_n}|/n \leqslant \eta$, and

$$\mathbf{P}(\widehat{S}_{n,\lambda_n} = S_{0,n}) \geqslant 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty. \tag{7}$$

The selection consistency is often met under certain sparsity conditions depending on the selection method (Zhao and Yu, 2006; Zhang, 2010). Take LASSO for example, the selection consistency property is guaranteed under $s_0 = O(n^{c_1})$ and $s_0 \log p = o(n^{c_3})$ for some $0 < c_1 < c_3 < 1$, along with irrepresentable condition and others.

THEOREM 2: *Given model (1) and assumptions (A1,A2,B3), consider the SPARES estimator* $\widehat{\beta}_{\text{SPARES}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^T$ *as defined in (4). For each* $j$, *there exist random variables*

$Z_j^0, \Delta_j$, *such that as* $n, B \to \infty$,

$$\sqrt{n}(\widehat{\beta}_j - \beta_j^0) = Z_j^0 + \Delta_j, \quad Z_j^0/\sigma_j \to N(0,1), \quad \Delta_j = o_p(1), \tag{8}$$

*where* $\sigma_j^2 = \sigma^2\left(\mathbf{\Sigma}^{-1}_{S_{0,n}\cup j}\right)_{jj}$ *is bounded.*

The proof is presented in the Web Appendix along with some useful lemmas. The difficulties in deriving the theoretical properties of the SPARES estimator arise primarily from the randomness of $S^b$'s, the selected subsets of variables from subsamples of the original data. It is unclear whether a standard bootstrap theorem can be applied to such random sets since the uniform control that one obtains under Donsker-type conditions in empirical process theory is absent. Consequently, assumptions weaker than selection consistency are not effective in controlling the randomness of the $S^b$'s. Meanwhile our simulations suggest the validity of SPARES when only (A3) holds instead of (B3). Under assumption (B3), the asymptotic variance of ours converges to the best variance of an unbiased estimator of $\beta_j^0$ under the reduced model

$$\boldsymbol{Y} = \boldsymbol{X}_{S_0\cup j}\beta^0_{S_0\cup j} + \varepsilon. \tag{9}$$

Such bound is smaller than the semiparametric information bound that involves all $p$ covariates (Belloni et al., 2014; Van de Geer et al., 2014). Nevertheless the sets of conditions for the mentioned works and ours are quite different that they might not be directly comparable.

## 4. Inference by SPARES

### 4.1 *Estimator of Standard Errors*

As shown in Theorem (2), $\widehat{\beta}_j$ converges to a normal distribution whose variance depends on the unknown active set $S_{0,n}$. We propose an implementable approach to estimating the standard error of $\widehat{\beta}_j$ using Theorem 1 of Efron (2014), see also Wager et al. (2014) and Theorem 9 of Wager and Athey (2018). We denote the estimator as $\widehat{\text{se}}_j^B$. For the $b^{th}$ bootstrap data, $D_1^b$, we re-write the index set as $I_1^b = (i_{b1}, i_{b2}, ..., i_{n/2})$. For $i = 1, 2, ..., n$

define $I_{bi} = \#\{i_{bk} = i\}$, the number of times that the $i^{th}$ observation appears in the $b^{th}$ re-sample. The vector $I^b = (I_{b1}, I_{b2}, ..., I_{bn})$ then follows a multinomial distribution with $n/2$ draws on $n$ outcomes each having probability $1/n$, whose mean vector and covariance matrix are

$$I^b \sim \left(\frac{1}{2}\mathbf{1}_n, \frac{1}{2}\mathbf{I}_n - \frac{1}{2n}\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right) \tag{10}$$

where $\mathbf{1}_n$ the (column) vector of $n$ 1's and $\mathbf{I}_n$ the $n \times n$ identity matrix. The nonparametric delta method estimator of the standard error is then given by:

$$\widehat{\mathrm{se}}_j^B = \left(\sum_{i=1}^{n} \widehat{\mathrm{cov}}_{ij}^2\right)^{1/2}, \tag{11}$$

where

$$\widehat{\mathrm{cov}}_{ij} = \sum_{b=1}^{B}(I_{bi} - \overline{I_{\cdot i}})(\widehat{\beta}_j^b - \widehat{\beta}_j)/B \tag{12}$$

is the bootstrap covariance between $I_{bi}$ and $\widehat{\beta}_j^b$, and $\overline{I_{\cdot i}} = \sum_{b=1}^{B} I_{bi}/B$.

As emphasized in Efron (2014), the merit of smoothing the SPARE estimator is to convert a "jumpy" selection-based estimator $\widehat{\beta}^b$ into a smooth version of $\widehat{\beta}$. It is pointed out in Wager et al. (2014) that the nonparametric delta method standard error estimator tends to be biased upwards when the number of bootstraps is small. They proposed an alternative bias-corrected version of (11):

$$\widehat{\mathrm{se}}_U^B = \left\{(\widehat{\mathrm{se}}^B)^2 - \frac{n}{2B^2}\sum_{b=1}^{B}(\widehat{\beta}^b - \widehat{\beta})^2\right\}^{1/2} \tag{13}$$

Note that (13) converges to (11) as $B \to \infty$. The original version (11) would require $B = O(n^{1.5})$ to reduce Monte Carlo noise down to the level of sampling noise, while (13) only requires $B = O(n)$. Moreover, our experience shows that the unbiased version does converge to the empirical standard error faster than the original one.

### 4.2 *Confidence Intervals and P-values*

Following previous discussion, the asymptotic $1-\alpha$ confidence interval for each $\beta_j^0$ is given by

$$\left(\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}_j^B \ , \ \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}_j^B\right), \tag{14}$$

where $\Phi^{-1}$ is the inverse CDF of the standard normal distribution. The p-value of testing $H_0 : \beta_j = 0$ is

$$p_j = 2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j|/\widehat{\mathrm{se}}_j^B\right)\right\}. \tag{15}$$

## 5. Extension of SPARES to a Subvector $\beta^{(1)}$ with a Fixed Dimension

It is natural to extend our procedure to a subvector $\beta^{(1)}$ of $\beta^0$ with a fixed dimension $p_1 \geqslant 2$. Without loss of generality, assume that $\beta^{(1)} = \beta^0_{S^{(1)}} = (\beta_1^0, \beta_2^0, .., \beta_{p_1}^0)^{\mathrm{T}}$ with $|S^{(1)}| = p_1$. Accordingly, we modify the SPARE estimator in (2) to be

$$\widehat{\beta}^b_{S^{(1)}} = \left\{ ({\boldsymbol{X}^b_{S^b \cup S^{(1)}}}^{\mathrm{T}} \boldsymbol{X}^b_{S^b \cup S^{(1)}})^{-1} {\boldsymbol{X}^b_{S^b \cup S^{(1)}}}^{\mathrm{T}} \boldsymbol{Y}^b \right\}_{S^{(1)}}, \tag{16}$$

which gives a corresponding SPARES estimator for $\beta^{(1)}$:

$$\widehat{\beta}^{(1)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}^b_{S^{(1)}}. \tag{17}$$

The corresponding extension of Theorem 2 is stated below.

THEOREM 3: *Consider model (1) under assumptions (A1,A2,B3), and a fixed finite subset $S^{(1)} \subset \{1, 2, .., p\}$ with $|S^{(1)}| = p_1$. Let $\widehat{\beta}^{(1)}$ be the SPARES estimator for $\beta^{(1)} = \beta^0_{S^{(1)}}$ as defined in (17). There exist random vectors $Z^{(1)}, \Delta^{(1)}$, such that as $n, B \to \infty$,*

$$\sqrt{n}(\widehat{\beta}^{(1)} - \beta^{(1)}) = Z^{(1)} + \Delta^{(1)}, \quad {\boldsymbol{\Sigma}^{(1)}}^{-1/2} Z^{(1)} \to N(0, \mathbf{I}_{p_1}), \quad \Delta^{(1)} = o_p(\mathbf{1}_{p_1}), \tag{18}$$

*and $\boldsymbol{\Sigma}^{(1)} = \sigma^2 \left(\boldsymbol{\Sigma}^{-1}_{S_{0,n} \cup S^{(1)}}\right)_{S^{(1)}}$ is positive definite.*

REMARK 3: There is also a direct extension of the one-dimensional nonparametric delta

method for estimating the variance-covariance matrix of $\widehat{\beta}^{(1)}$, $\widehat{\mathbf{\Sigma}}^{(1)} = \widehat{\mathrm{COV}}_{(1)}^{\mathrm{T}}\widehat{\mathrm{COV}}_{(1)}$, where

$$\widehat{\mathrm{COV}}_{(1)} = \left(\widehat{\mathrm{cov}}_1^{(1)}, \widehat{\mathrm{cov}}_2^{(1)}, .., \widehat{\mathrm{cov}}_n^{(1)}\right)^{\mathrm{T}} \tag{19}$$

$$\widehat{\mathrm{cov}}_i^{(1)} = \sum_{b=1}^{B}(I_{bi} - I_{\cdot i})(\widehat{\beta}^b_{S^{(1)}} - \widehat{\beta}^{(1)})/B. \tag{20}$$

The extension to a subvector $\beta^{(1)}$ with a fixed dimension allows us to derive confidence regions for a subset of variables of interest and test for contrasts of certain predictors.

## 6. Simulation Studies

We designed all simulation scenarios based on the linear model (1) with $\boldsymbol{X} = (X_1, ...X_p) = (\mathbf{x}_1^{\mathrm{T}}, ..., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathrm{T}}$, assuming $\mathbf{x}_i$'s i.i.d. $\sim N(\mathbf{0}_p, \mathbf{\Sigma}_{p\times p})$ and $\varepsilon_i$'s i.i.d. $\sim N(0, 1)$. A total of 200 simulated datasets were generated for each simulation configuration.

We first illustrate the advantage of using SPARES over one-time SPARE. We set sample size $n = 200$, number of predictors $p = 300$, and $s_0 = 3$ nonzero signals with $\mathbf{\Sigma}_{p\times p}$ being the identity matrix. As shown in Web Table (1), over 200 replications, the biases of both approaches are negligible on average, but the standard errors of SPARES are much smaller than those of one-time SPARE, which results in higher power and more accurate inferences. Thus we recommend SPARES in practice.

In subsection 6.1, we explore the performance of SPARES under various settings, including different correlation structures of $\boldsymbol{X}$, strong and weak signals strength, and stress tests with ultrahigh dimensionality. In subsection 6.2, we compare SPARES with two de-biased LASSO estimators, LASSO-Pro from Van de Geer et al. (2014) and SSLASSO from Javanmard and Montanari (2014).

### 6.1 *Performance of SPARES under various settings*

We will go over three examples, all of which assume the linear model (1) as truth, but with different parameters.

**Example 1.** Let sample size $n = 150$, number of predictors $p = 300$, number of nonzero signals $s_0 = 5$, and a fixed realization of $\beta^0$ where $S_{0,n} = \{66, 97, 145, 166, 173\}$ was a fixed realization of $s_0$ draws without replacement from $[p]$ and $\beta^0_{S_{0,n}} = (1, 0.6, -1, -0.6, 1)$. We examined three commonly used correlation structures: identity; first-order autoregressive (AR(1)) with $\rho = 0.5$; compound symmetry (CS) with $\rho = 0.5$. LASSO was used as the selection procedure $\mathcal{S}_\lambda$, while $\lambda$ was chosen by cross-validation. As summarized in Table (1), for both nonzero signals and noise variables, the bias of SPARES estimator was well controlled while the SE estimates were very close to the empirical ones. Consequently, the coverage probabilities of the 95% confidence intervals were at the nominal level. In addition, the variable selection frequency based on p-values of SPARES was higher for true signals and much lower for noise variables compared to selection by LASSO. Notice that for identity and AR(1) correlation structures, the selection frequencies of the true signals were uniformly close to 1, suggesting "sure screening" condition was met and thus the better coverage probabilities. Therefore the simulation result validates our claim that SPARES works under "sure screening" assumption.

**Example 2.** Let $n = 150,\ p = 500$, and

- Example 2.1: $s_0 = 15$, $\mathbf{\Sigma}_{p\times p} = \text{diag}(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10})$, where each $\mathbf{\Sigma}_k$ was $50 \times 50$ with an AR(1) correlation structure, $(\mathbf{\Sigma}_k)_{ij} = (0.1k - 0.1)^{|i-j|},\ k = 1, 2, .., 10$. The active set $S_{0,n}$ was a fixed realization of $s_0$ draws without replacement from $[p]$, and $\beta^0_{S_{0,n}}$ was a fixed realization of $s_0$ i.i.d. Uniform $U[0, 2]$ variables;
- Example 2.2: $s_0 = 20$, $\mathbf{\Sigma}_{p\times p} = \text{diag}(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10})$, where each $\mathbf{\Sigma}_k : (\mathbf{\Sigma}_k)_{ij} = (0.3)^{|i-j|}$. The non-zero signals are assigned effect sizes $\beta^0_{50k-45} = 0.2k,\ \beta^0_{50k-15} = -0.2k$ for $k = 1, 2, ..., 10$.

We applied SPARES with LASSO (10-fold cross validation to choose $\lambda$) as the model selection procedure, and reported the simulation averages of $\widehat{\beta}_{\text{SPARES}}$, along with confidence intervals, mean biases, coverage probabilities, and type I errors for testing $H_0 : \beta^0_j = 0$.

The results are summarized in Web Figures (1) and (2). For the true signals $j \in S_{0,n}$, the proposed method worked well regardless of the correlation, with negligible biases and close-to-nominal coverage probabilities. On the other hand, the biases for the estimates of noise variables were enlarged when they were highly correlated with non-zero signals. The estimated coverage probabilities and type I errors deviated more from the nominal level consequently. The type I error became negligible when the effect size was over 1. Coupled with an observation that the bias was larger for the noise variables that were correlated with moderate non-zero signals, our takeaway was that the magnitude of bias was a combination of selection errors as well as correlations with true signals.

**Example 3** serves as a "stress test" to illustrate how SPARES handle large datasets with a number of "weak signals". We let $n = 500$, $p = 1000,\ 5000$ and 10000, and $s_0 = 205$. Within the 205 non-zero signals, 5 are of sizes $0.2, 0.4, 0.6, 0.8, 1$, and the rest 200 are fixed random realizations from the uniform distribution $U[(-0.2, -0.1) \cup (0.1, 0.2)]$. The multivariate normal distribution with mean zero and the AR(1) correlation structure with $\rho = 0.5$ is applied to generate $\boldsymbol{X}$'s. As summarized in Table (2), the SPARES estimator remains nearly unbiased for both strong and weak signals. The coverage probabilities of strong signals are close to the nominal level 0.95, while those for weak and zero signals are above 0.9 on average. This demonstrates that SPARES is rather reliable and robust even for large datasets with a number of weak signals.

### 6.2 *Comparisons with De-biased LASSO Estimators*

We compared SPARES with different versions of de-biased LASSO estimators in Example 4, where the active set $S_{0,n} \subset \{1, 2, .., p\}$ was a fixed random realization with size $|S_{0,n}| = 5$, and $\beta^0_{S_{0,n}}$ was a fixed realization of 5 i.i.d. random variables from uniform $U[0.5, 2]$. The size of the active set is reduced to 5 for clearer comparison and display of the result. Three correlation structures are considered for completeness:

- Example 4.1: Identity $\mathbf{\Sigma}_{p\times p} = \mathbf{I}_{p\times p}$;
- Example 4.2: AR(1) $\mathbf{\Sigma}_{p\times p} : (\mathbf{\Sigma})_{jk} = (0.8)^{|j-k|}$;
- Example 4.3: Compound symmetry $\mathbf{\Sigma}_{p\times p} : (\mathbf{\Sigma})_{jk} = 0.5$.

The estimated biases and coverage probabilities were shown in Table (3) and Web Figure (3), where LASSO-Pro was proposed in Van de Geer et al. (2014) and SSLASSO was from Javanmard and Montanari (2014).

Across the board, SPARES gave less biased point estimates for the true signals, and provided reliable confidence intervals around the nominal level for both true signals and noise variables. In contrast, both LASSO-Pro and SSLASSO had visible discrepancies between the true signals and noise variables. While LASSO-Pro had lower-than-nominal level coverages for the true signals, it performed even worse in Example 4.1, probably due to the fact that the node-wise LASSO was not ideal when estimating the precision matrix when $\mathbf{\Sigma}_{p\times p}$ was an identity matrix. As far as SSLASSO was concerned, the confidence intervals for the noise variables were too conservative, while the coverages for the true signals in Example 4.2 were considerably low.

In summary, the performance of SPARES aligned well with the theoretical expectations, especially for the active set $S_{0,n}$. We did observe, however, some false-positives when the noise variables were highly correlated with those in the active set. Nevertheless, compared with the de-biased LASSO methods, SPARES showed substantial improvement by providing less biased estimates with more accurate coverage probabilities close to the nominal level.

## 7. Data Examples

### 7.1 *Riboflavin Production Data*

We applied our method to analyze a dataset on riboflavin (vitamin $B_2$) production by bacillus subtilis, made public by Bühlmann et al. (2014) and analyzed by Meinshausen et al. (2009),

Bühlmann et al. (2014), Van de Geer et al. (2014) and Javanmard and Montanari (2014). The data contained $n = 71$ samples and $p = 4088$ covariates, measuring the logarithm of the expression levels of 4088 genes. The response variable was the logarithm of the riboflavin production rate.

We related the response to the gene expressions using the linear model (1). We checked the collinearity among the genes, and their pairwise correlations are plotted in the Web Figure (4). We further normalized the genes so that their effect sizes are comparable. The LASSO was used as the variable selection method, and we let $B = 1000$ be the number of re-samples. Assisted by the LASSO selection, we derive the SPARES estimator $\widehat{\beta}$, the standard error estimates as in (11), and the p-values as in (15). With a standard Bonferroni correction to adjust FWER to the 5% significance level, we identified four genes that were significantly associated with the response, namely YCKE_at, XHLA_at, YXLD_at, and YDAR_at. If the FWER were set at 10%, one more gene, YCGN_at, would be included. The confidence intervals for the top 5 genes are displayed on the right panel of Web Figure (5), with the point estimates shown in Table (4). By contrast, the results from other methods were less informative. For example, with a 5% FWER, the multisample-splitting method proposed in Meinshausen et al. (2009) identified YXLD_at, Van de Geer et al. (2014) claimed none, and Javanmard and Montanari (2014) only detected YXLD_at and YXLE_at, which are highly correlated themselves.

Our results had biological interpretations that are confirmed by the literature. It was reported that XHLA_at was involved in cell lysis upon induction of PbsX (Kunst et al., 1997), increasing the capability to produce recombinant extracellular digestive enzymes that results in riboflavin production (7.04 in Mander and Liu (2010)). YCKE_at, formally named as bglC, was also responsible for the production of certain enzyme, Aryl-phospho-beta-D-glucosidase,

and had extracellular protein secretory functions (Schallmey et al., 2004). YXLD_at, together with YXLE_at, was important for negative regulation of sigma Y activity (Tojo et al., 2003).

### 7.2 *Multiple Myeloma Genomic Data*

We analyzed a cancer genomic data with $n = 163$ multiple myeloma patients. Our interest lay in detecting the association between the $\beta$-2 microglobulin (B2M) and gene expressions. B2M is a small membrane protein produced by malignant myeloma cells, indicating the severity of disease. Identifying genes that are related to B2M is clinically important as it helps construct molecular prognostic tools for early diagnosis of disease.

We first used KEGG (Carlson, 2015) to identify gene pathways that are related to cancer development and progression, as well as some identified upstream genes that may regulate B2M. In total, there were $p = 789$ unique probes belonging to these pathways. We took the logarithm transformation for both the B2M test value and the gene expressions as our response and predictors for model (1). We applied SPARES with LASSO as the selection method, and $B = 500$ re-samples were drawn for smoothing.

Our method offers additional biological insight compared to the other methods. As shown in Table (5), it identified two significant probes at 5% FWER after the Bonferroni correction, namely 204171_at (RPS6KB1) and 202076_at (BIRC2). In contrast, the two de-biased LASSO estimators identified no significant probes. Both detected genes are highly associated with malignant tumor cells: RPS6KB1, member of the ribosomal protein S6 kinase (RPS6K) family, altercation/mutation has been related to numerous types of cancer including breast cancer, colon cancer, non-small-cell lung cancer, and prostate cancer (Sinclair et al., 2003; Van der Hage et al., 2004; Slattery et al., 2011; Zhang et al., 2013; Cai et al., 2015); BIRC2, whose encoded protein is a member of inhibitors of apoptotic proteins (IAPs) that inhibits apoptosis by binding to tumor necrosis factor receptor-associated factors TRAF1 and TRAF2

(Saleem et al., 2013), has been related to lung cancer and lymphoma (Wang et al., 2010; Rahal et al., 2014).

## 8. Conclusion

We have proposed a new framework of estimation and inference for the high-dimensional linear models (1), and shown the proposed SPARES estimator is asymptotically unbiased and normal, giving accurate and reliable component-wise inferences. The key improvement, compared to the existing works, lies in these aspects. SPARES converts the high-dimensional problem of estimating the $p$-vector $\beta^0$ to the low dimensional case by Selection-assisted Partial Regression. Thus we avoid the curse of dimensionality on estimation and inference. SPARES is applicable to general selection methods including LASSO, SCAD, screening, boosting, and etc., as long as they possess the desired selection consistency property, which is likely to be loosened to sure screening property in practice as suggested in the extensive simulation study. SPARES is not sensitive to the tuning parameter $\lambda$ in $\mathcal{S}_\lambda$, since it is not directly used for estimation, but only involved in the selection. Hence, our method has minimal requirements on extra model parameters and is almost robust toward selection of tuning parameters. This framework can be naturally extended to other non-linear regression models, such as generalized linear model and Cox model, through two general steps. First, we perform data-splitting on the original data, and then do selection on one half of the data followed by fitting low-dimensional model on the other half of the data using partial regression; Second, we repeat the first step many times and average over all estimates to form a smoothed estimate. We will report this work elsewhere.

Acknowledgements

The authors thank the Editor, the AE and two referees for their comments and suggestions that helped improve the manuscript. The work is supported by grants from the NIH, NSF and NSFC.

References

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81,** 608–650.

Belloni, A., Chernozhukov, V., and Wei, Y. (2013). Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41,** 802–837.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application* **1,** 255–278.

Cai, C., Chen, Q.-B., Han, Z.-D., Zhang, Y.-Q., He, H.-C., Chen, J.-H., et al. (2015). Mir-195 inhibits tumor progression by targeting rps6kb1 in human prostate cancer. *Clinical Cancer Research* **21,** 4922–4934.

Carlson, M. (2015). *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2).* R package version 3.2.2.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109,** 991–1007.

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. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* **38,** 3567–3604.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15,** 2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics* **46,** 2593–2622.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., et al. (1997). The complete genome sequence of the gram-positive bacterium bacillus subtilis. *Nature* **390,** 249–256.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44,** 907–927.

Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pages 136–144.

Mander, L. and Liu, H.-W. (2010). *Comprehensive Natural Products II: Chemistry and Biology*, volume 1. Elsevier.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104,** 1671–1681.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45,** 158–195.

Rahal, R., Frick, M., Romero, R., Korn, J. M., Kridel, R., Chan, F. C., et al. (2014).

Pharmacological and genomic profiling identifies nf-$\kappa$B-targeted treatment strategies for mantle cell lymphoma. *Nature Medicine* **20,** 87–92.

Saleem, M., Qadir, M. I., Perveen, N., Ahmad, B., Saleem, U., and Irshad, T. (2013). Inhibitors of apoptotic proteins: new targets for anticancer therapy. *Chemical Biology & Drug Design* **82,** 243–251.

Schallmey, M., Singh, A., and Ward, O. P. (2004). Developments in the use of bacillus species for industrial production. *Canadian Journal of Microbiology* **50,** 1–17.

Sinclair, C. S., Rowley, M., Naderi, A., and Couch, F. J. (2003). The 17q23 amplicon and breast cancer. *Breast Cancer Research and Treatment* **78,** 313–322.

Slattery, M. L., Lundgreen, A., Herrick, J. S., and Wolff, R. K. (2011). Genetic variation in rps6ka1, rps6ka2, rps6kb1, rps6kb2, and pdk1 and risk of colon or rectal cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **706,** 13–20.

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

Tojo, S., Matsunaga, M., Matsumoto, T., Kang, C.-M., Yamaguchi, H., Asai, K., et al. (2003). Organization and expression of the bacillus subtilis sigy operon. *Journal of Biochemistry* **134,** 935–946.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42,** 1166–1202.

Van der Hage, J., van den Broek, L., Legrand, C., Clahsen, P., Bosch, C., Robanus-Maandag, E., et al. (2004). Overexpression of p70 s6 kinase protein is associated with increased risk of locoregional recurrence in node-negative premenopausal early breast cancer patients. *British Journal of Cancer* **90,** 1543–1550.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects

using random forests. *Journal of the American Statistical Association* **113,** 1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* **15,** 1625–1651.

Wang, Y., Dong, Q., Zhang, Q., Li, Z., Wang, E., and Qiu, X. (2010). Overexpression of yes-associated protein contributes to progression and poor prognosis of non-small-cell lung cancer. *Cancer Science* **101,** 1279–1285.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37,** 2178–2201.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76,** 217–242.

Zhang, Y., Ni, H.-J., and Cheng, D.-Y. (2013). Prognostic value of phosphorylated mtor/rps6kb1 in non-small cell lung cancer. *Asian Pacific Journal of Cancer Prevention* **14,** 3725–3728.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article, including Web Appendices, Tables, and Figures referenced in Sections 2-7.

Software R implementation of SPARES is available on-line at https://github.com/feizhe/SPARES, along with the simulation examples.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

*Received March* 2018. *Revised Nov* 2018. *Accepted Dec* 2018.

Table 1: Performance of SPARES under simulation example 1 with three correlation structures: Identity, AR(1) and Compound Symmetry (CS). The last column "-" represents the averages for all noise variables. **Freq $\mathcal{S}_\lambda$** is the selection frequency by LASSO; **Freq SPARES** is the selection frequency by p values of SPARES with 0.1 FDR control; **Empirical SE** is the empirical standard error.

| | Index $j$ | 66 | 97 | 145 | 166 | 173 | - |
|---|---|---|---|---|---|---|---|
| | $\beta_j^0$ | 1 | 0.6 | -1 | -0.6 | 1 | 0 |
| Identity | Bias ($\times 10^{-3}$) | 16 | -1 | -2 | 2 | 7 | -1 |
| | Average $\widehat{\text{se}}_j^B$ | 0.110 | 0.111 | 0.109 | 0.111 | 0.110 | 0.111 |
| | Empirical SE | 0.117 | 0.109 | 0.104 | 0.113 | 0.124 | 0.109 |
| | Cov Prob (%) | 91.5 | 94.0 | 95.0 | 96.0 | 91.5 | 94.8 |
| | Freq $\mathcal{S}_\lambda$ | 1 | 0.956 | 1 | 0.965 | 1 | 0.059 |
| | Freq SPARES | 1 | 0.97 | 1 | 0.99 | 1 | 0.003 |
| AR(1) | Bias ($\times 10^{-3}$) | -6 | 2 | 7 | 10 | -1 | 0 |
| | Average $\widehat{\text{se}}_j^B$ | 0.115 | 0.116 | 0.114 | 0.115 | 0.116 | 0.115 |
| | Empirical SE | 0.125 | 0.108 | 0.114 | 0.120 | 0.108 | 0.114 |
| | Cov Prob (%) | 93.5 | 96.0 | 95.0 | 92.5 | 96.5 | 94.5 |
| | Freq $\mathcal{S}_\lambda$ | 0.998 | 0.938 | 1.000 | 0.929 | 1.000 | 0.046 |
| | Freq SPARES | 1 | 0.925 | 1 | 0.905 | 1 | 0.001 |
| CS | Bias ($\times 10^{-3}$) | -12 | -30 | 6 | 7 | -14 | -7 |
| | Average $\widehat{\text{se}}_j^B$ | 0.151 | 0.149 | 0.152 | 0.150 | 0.150 | 0.154 |
| | Empirical SE | 0.165 | 0.161 | 0.168 | 0.162 | 0.163 | 0.154 |
| | Cov Prob (%) | 92.5 | 91.5 | 89.4 | 92.0 | 92.0 | 94.5 |
| | Freq $\mathcal{S}_\lambda$ | 0.986 | 0.742 | 0.958 | 0.651 | 0.988 | 0.045 |
| | Freq SPARES | 1 | 0.775 | 1 | 0.795 | 1 | 0.005 |

Table 2: Performance of SPARES under simulation Example 3. Tables from top to bottom correspond to $p = 1000,\ 5000$ and 10000. Last two columns are averages over small and zero signals.

| Index | 36 | 272 | 376 | 568 | 915 | Small | 0's |
|---|---|---|---|---|---|---|---|
| $\beta^0$ | 0.200 | 0.400 | 0.600 | 0.800 | 1.000 | | 0.000 |
| | | | | $p = 1000$ | | | |
| Bias | 0.013 | -0.006 | 0.014 | -0.002 | -0.014 | 0.005 | 0.004 |
| Avg SE | 0.093 | 0.093 | 0.093 | 0.093 | 0.093 | 0.093 | 0.093 |
| Emp SE | 0.099 | 0.098 | 0.098 | 0.093 | 0.097 | 0.094 | 0.094 |
| Cov Prob | 0.960 | 0.920 | 0.930 | 0.930 | 0.940 | 0.907 | 0.908 |
| Sel freq | 0.045 | 0.418 | 0.930 | 1.000 | 1.000 | 0.021 | 0.002 |
| | | | | $p = 5000$ | | | |
| Bias | -0.005 | 0.009 | 0.010 | 0.003 | 0.004 | 0.004 | 0.000 |
| Avg SE | 0.093 | 0.093 | 0.095 | 0.094 | 0.094 | 0.094 | 0.094 |
| Emp SE | 0.092 | 0.096 | 0.098 | 0.099 | 0.112 | 0.095 | 0.096 |
| Cov Prob | 0.960 | 0.930 | 0.960 | 0.910 | 0.920 | 0.905 | 0.935 |
| Sel freq | 0.022 | 0.390 | 0.906 | 0.999 | 1.000 | 0.015 | 0.001 |
| | | | | $p = 10000$ | | | |
| Bias | -0.003 | 0.003 | 0.006 | 0.008 | -0.025 | 0.005 | 0.000 |
| Avg SE | 0.094 | 0.094 | 0.094 | 0.095 | 0.094 | 0.095 | 0.095 |
| Emp SE | 0.094 | 0.096 | 0.101 | 0.103 | 0.093 | 0.096 | 0.097 |
| Cov Prob | 0.950 | 0.940 | 0.930 | 0.930 | 0.950 | 0.902 | 0.939 |
| Sel freq | 0.015 | 0.313 | 0.860 | 0.996 | 1.000 | 0.012 | 0.000 |

Table 3: Comparisons of SPARES with LASSO-Pro and SSLASSO under Example 4. The rows consist of 5 true signals and the average of zero signals. In each cell, top number is for SPARES; middle number is for LASSO-Pro; lower number is for SSLASSO.

| | | Example 4.1 | | Example 4.2 | | Example 4.3 | |
|---|---|---|---|---|---|---|---|
| Index | $\beta_j^0$ | Bias ($\times 10^{-3}$) | Cov Prob (%) | Bias ($\times 10^{-3}$) | Cov Prob (%) | Bias ($\times 10^{-3}$) | Cov Prob (%) |
| | | -1.77 | 90.5 | 10.43 | 92.5 | -0.35 | 96.5 |
| | | -81.78 | 70.5 | -44.09 | 86 | -38.43 | 92.5 |
| 78 | 1.07 | -79.33 | 90.5 | -101.95 | 84.5 | -113.72 | 92.5 |
| | | -1.04 | 96.5 | 9.70 | 92 | 2.44 | 95 |
| | | -80.28 | 76 | -44.54 | 87 | -32.42 | 89 |
| 102 | 1.04 | -77.72 | 93.5 | -99.66 | 82 | -105.60 | 92 |
| | | -1.62 | 94 | 15.58 | 93.5 | -4.67 | 96.5 |
| | | -89.43 | 71.5 | -47.57 | 88.5 | -40.39 | 91.5 |
| 242 | 1.19 | -88.69 | 87.5 | -104.25 | 84 | -115.51 | 92 |
| | | -0.14 | 94 | 2.98 | 96.5 | 2.01 | 95 |
| | | -75.87 | 81 | -41.40 | 88 | -30.61 | 91 |
| 359 | 1.43 | -80.91 | 94 | -98.14 | 85 | -107.5 | 89 |
| | | -3.57 | 95.5 | 0.54 | 93 | 5.88 | 91.5 |
| | | -84.86 | 75 | -60.80 | 88 | -24.20 | 86.5 |
| 380 | 0.62 | -85.73 | 89.5 | -111.11 | 81.5 | -99.26 | 90.5 |
| | | -0.46 | 95 | 0.65 | 94.82 | 3.26 | 95.16 |
| | | -0.40 | 97 | 3.16 | 96.46 | 5.24 | 96.34 |
| - | 0 | -0.27 | 99.5 | 4.15 | 99.69 | 26.88 | 99.94 |

Table 4: Analysis of the riboflavin genomic data. $\widehat{\beta}$ is the SPARES estimator; $p$-values are adjusted by Bonferroni correction (multiplied by $p$). The top 10 and bottom 10 most/least significant genes are tabulated.

| Gene | $\widehat{\beta}$ | SE | Adjusted $p$-value |
|---|---|---|---|
| YCKE_at | 0.37 | 0.06 | $< 0.001$ |
| XHLA_at | 0.48 | 0.09 | $< 0.001$ |
| YXLD_at | -0.53 | 0.11 | 0.01 |
| YDAR_at | -0.28 | 0.06 | 0.01 |
| YCGN_at | -0.31 | 0.07 | 0.09 |
| RPLJ_at | -0.26 | 0.06 | 0.10 |
| YQIZ_at | -0.25 | 0.06 | 0.13 |
| YCDH_at | -0.27 | 0.07 | 0.15 |
| SPOIISA_at | 0.25 | 0.06 | 0.35 |
| YRPE_at | -0.25 | 0.07 | 0.63 |
| ... | | | |
| YXAL_at | $-2 \times 10^{-4}$ | 0.09 | 1 |
| XPT_at | $-1.6 \times 10^{-4}$ | 0.07 | 1 |
| YOZG_at | $-2.9 \times 10^{-4}$ | 0.14 | 1 |
| YOJB_at | $1.7 \times 10^{-4}$ | 0.10 | 1 |
| YBCL_at | $-1.8 \times 10^{-4}$ | 0.11 | 1 |
| YJAX_at | $1.3 \times 10^{-4}$ | 0.09 | 1 |
| YOSE_at | $1.1 \times 10^{-4}$ | 0.11 | 1 |
| YUNA_at | $4.9 \times 10^{-5}$ | 0.07 | 1 |
| YISO_at | $1.7 \times 10^{-5}$ | 0.08 | 1 |

Table 5: Analysis of the Multiple Myeloma genomic data. The top 6 and bottom 6 most/least significant genes are tabulated.

| Gene | $\widehat{\beta}$ | SE | Adjusted $p$ |
|---|---|---|---|
| 204171_at (RPS6KB1) | -0.20 | 0.042 | 0.002 |
| 202076_at (BIRC2) | -0.17 | 0.041 | 0.037 |
| 220414_at | -0.20 | 0.05 | 0.14 |
| 220394_at | -0.18 | 0.05 | 0.59 |
| 206493_at | -0.19 | 0.06 | 0.63 |
| 209878_s_at | -0.17 | 0.05 | 0.69 |
| ... | | | |
| 207924_x_at | $5 \times 10^{-4}$ | 0.07 | 1 |
| 205289_at | $-4.4 \times 10^{-4}$ | 0.06 | 1 |
| 203591_s_at | $4.7 \times 10^{-4}$ | 0.07 | 1 |
| 224229_s_at | $2.4 \times 10^{-4}$ | 0.06 | 1 |
| 217576_x_at | $2.5 \times 10^{-4}$ | 0.07 | 1 |
| 201656_at | $2.5 \times 10^{-4}$ | 0.08 | 1 |