JUSTC

Article

http://justc.ustc.edu.cn

Received: March 11, 2024; Accepted: June 06, 2024

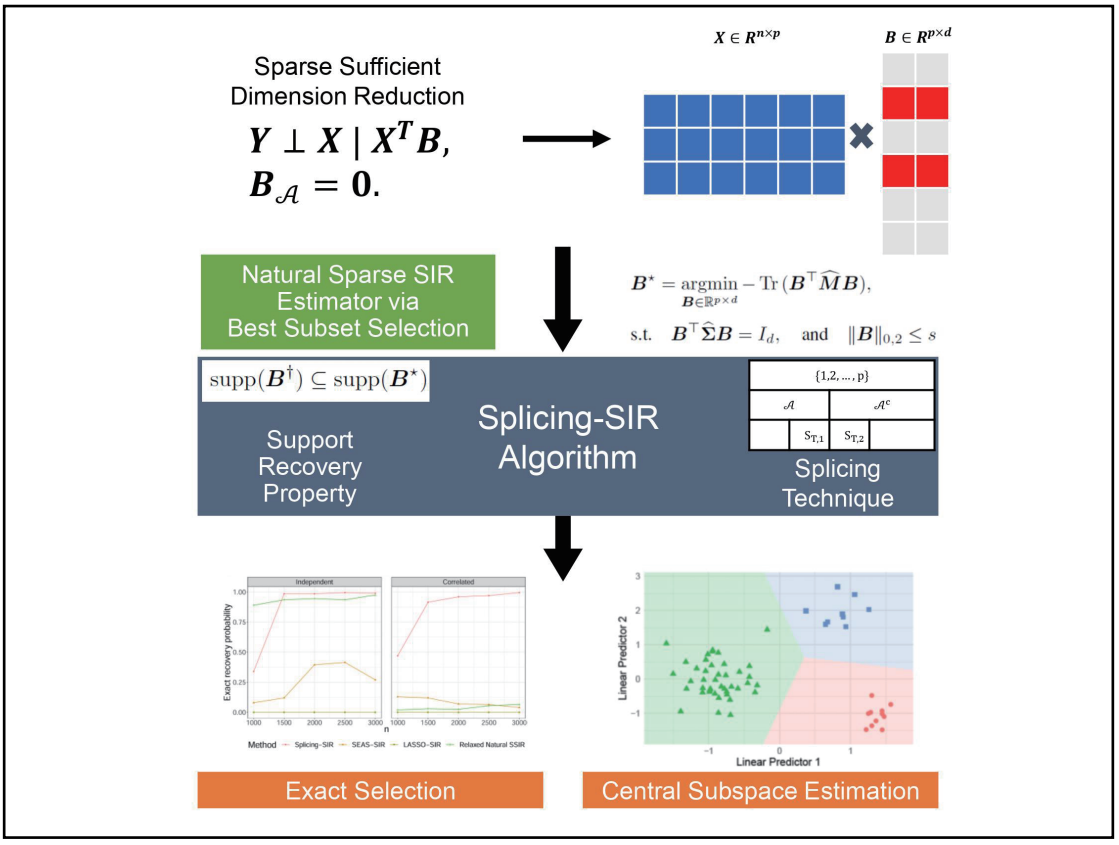
A splicing algorithm for best subset selection in sliced inverse regression

Borui Tang[1], Jin Zhu[2], Tingyin Wang[1], and Junxian Zhu[3] ✉

[1]*International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China;*
[2]*Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom;*
[3]*Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117549, Singapore*

✉Correspondence: Junxian Zhu, E-mail: junxian@nus.edu.sg

Graphical abstract



The splicing-SIR algorithm with methodological insights and empirical demonstration.

Public summary

- We propose a novel splicing algorithm to solve the natural sparse sliced inverse regression estimator.
- This algorithm directly and simultaneously tackles the sparsity and orthogonal constraints by iteratively approximating the optimal conditions.
- Empirically, it is fast, capable of exactly recovering the best subset, accurate in central subspace estimation, and robust against design dependence.

Citation: Tang B R, Zhu J, Wang T Y, et al. A splicing algorithm for best subset selection in sliced inverse regression. *JUSTC*, **2025**, 55(5): 0503. CSTR: 32290.14.JUSTC-2024-0034. DOI: 10.52396/JUSTC-2024-0034

JUSTC

Article

http://justc.ustc.edu.cn

Received: March 11, 2024; Accepted: June 06, 2024

A splicing algorithm for best subset selection in sliced inverse regression

Borui Tang[1], Jin Zhu[2], Tingyin Wang[1], and Junxian Zhu[3] ✉

[1]*International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China;*
[2]*Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, United Kingdom;*
[3]*Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117549, Singapore*

✉Correspondence: Junxian Zhu, E-mail: junxian@nus.edu.sg

Cite This: *JUSTC*, **2025**, 55(5): 0503 (13pp) Read Online

Abstract: In this paper, we examine the problem of sliced inverse regression (SIR), a widely used method for sufficient dimension reduction (SDR). It was designed to find reduced-dimensional versions of multivariate predictors by replacing them with a minimally adequate collection of their linear combinations without loss of information. Recently, regularization methods have been proposed in SIR to incorporate a sparse structure of predictors for better interpretability. However, existing methods consider convex relaxation to bypass the sparsity constraint, which may not lead to the best subset, and particularly tends to include irrelevant variables when predictors are correlated. In this paper, we approach sparse SIR as a nonconvex optimization problem and directly tackle the sparsity constraint by establishing the optimal conditions and iteratively solving them via the splicing technique. Without employing convex relaxation on the sparsity constraint and the orthogonal constraint, our algorithm exhibits superior empirical merits, as evidenced by extensive numerical studies. Computationally, our algorithm is much faster than the relaxed approach for the natural sparse SIR estimator. Statistically, our algorithm surpasses existing methods in terms of accuracy for central subspace estimation and best subset selection and sustains high performance even with correlated predictors.

Keywords: splicing technique; best subset selection; sliced inverse regression; nonconvex optimization; sparsity constraint; optimal conditions

CLC number: O212.1 **Document code:** A

2020 Mathematics Subject Classification: 62H12

1 Introduction

The rapid advancement of data collection technology across various fields has led to an evident increase in the dimensionality of predictors and created challenges for traditional multivariate modeling. However, for most of the scenarios, only a small collection of linear combinations of predictors may contribute to the response. To uncover the underlying low-dimensional patterns, researchers have proposed a statistical framework called sufficient dimension reduction (SDR). It refers to the statistical tool that reduces the dimension of the predictors by replacing the original predictors with a minimal set of their linear combinations without loss of information. We refer to Refs. [1, 2] for early works on SDR, and Ref. [3] for a comprehensive overview. We mathematically formulate the SDR problem as follows. Let $\boldsymbol{X} = (X_1, \cdots, X_p)^\top$ be the predictor and Y be the scalar response. The aim of SDR is to search for $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with $d \le p$, such that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^\top \boldsymbol{X}, \tag{1}$$

that is, Y is independent of $\boldsymbol{X}$ conditioning on $\boldsymbol{\beta}^\top \boldsymbol{X}$. The column space of $\boldsymbol{\beta}$, Span($\boldsymbol{\beta}$), is called a dimension reduction space. Under mild conditions, Cook[4] justified that the intersection of all dimension reduction spaces is itself a dimension reduction space, and we refer to it as the central subspace. Among the various methods proposed to estimate the central subspace, sliced inverse regression (SIR) is notably popular and commonly used. Since it was proposed in Ref. [1], it has attracted a great deal of attention[5–7].

SIR employs all predictors to construct the central subspace, which often poses challenges in interpreting reduced-dimensional predictors. To this end, it is commonly assumed that only a subset of predictors contributes to the central subspace, giving rise to the sparse SIR problem[8]. Specifically, researchers search for $\boldsymbol{\beta}$ with a row-wise sparsity structure, that is, $\boldsymbol{\beta}_{\mathcal{A}} = \boldsymbol{0}$ for an index set $\mathcal{A} \in \{1, 2, \cdots, p\}$. We refer to $\mathcal{A}$ as the active set, and its complementary set in $\{1, 2, \cdots, p\}$ as the inactive set. This sparsity structure is closely related to model-free variable selection[9], which searches for the smallest subset $\mathcal{A}$ such that

$$Y \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{A}^c} | \boldsymbol{X}_{\mathcal{A}}. \tag{2}$$

The existence and uniqueness of the smallest $\mathcal{A}$ have been established in Ref. [10]. As noted in Ref. [11], (2) is equivalent to $\boldsymbol{\beta}_{\mathcal{A}} = \boldsymbol{0}$ in (1). This equivalency implies that by searching for $\boldsymbol{\beta}$ with a row-wise sparsity structure, we can simultan-

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

eously achieve central subspace estimation and best subset selection.

To address the sparse SIR problem, various regularization methods have been developed in the literature. In the low-dimensional setting, the row-wise sparsity can be encouraged by a coordinate-independent penalty[12], a LASSO penalty[13], or a SCAD max penalty[14]. In the high-dimensional setting, Lin et al.[15] introduced a Lasso-type approach for sequentially estimating sparse SIR directions. Concurrently, Tan et al.[16] transformed the sparse SIR optimization challenge into a generalized Rayleigh quotient problem, while Tan et al.[17] developed a convex formulation to extend the SIR method to high dimensions. Furthermore, Lin et al.[18] and Tan et al.[19] studied the theoretical limits and optimal rates of sparse SIR. More recently, Zeng et al.[20] connected the double penalization technique[21, 22] with the SDR framework, leading to a unified approach known as Subspace Estimation with Automatic Dimension and Variable Selection (SEAS). They formulated the general SDR problem as quadratic convex optimization, adopting a nuclear norm penalty for dimension selection as well as a group Lasso penalty for coordinate-independent variable selection.

Regularization methods are adopted in the above works to obtain the sparsity structure. Known as the convex relaxation of the best subset selection problem, regularization is widely used but has limitations. In particular, these methods may not exactly lead to the best subset, sometimes including irrelevant variables, especially when the design variables are correlated. This issue has been highlighted in Ref. [23]. Despite the limitations of convex relaxation methods, the best subset selection problem in sparse SIR has not been thoroughly studied because, with the sparsity constraint, this problem is computationally intractable. To overcome this limitation, we develop a novel algorithm for sparse SIR in this article to address this computational challenge and further advance the field. Named splicing-SIR, our approach integrates the sparsity constraint into the algorithm, offering a novel solution to the limitations of existing methods. Motivated by the success of splicing iterations across different models and scenarios, including linear regression[24], reduced-rank regression[25], generalized linear models[26], and single index models[27], we develop a splicing algorithm for the nonconvex sparsity-constrained optimization problem.

Our established method contributes to the literature in two aspects. Methodologically, we propose a novel sparse SIR algorithm. After establishing and investigating the optimal conditions of the optimization problem, we directly and simultaneously tackle the sparsity constraint and the orthogonal constraint by iteratively approximating the optimal conditions utilizing the technique of splicing iterations. Empirically, our proposed algorithm has superior empirical performance: it is computationally efficient, accurate in central subspace estimation, robust against design dependence, and able to exactly recover the best subset with a sparse model more parsimonious than other state-of-the-art methods.

The rest of the article is organized as follows. In Section 2, we introduce the methodology of algorithmic details of splicing-SIR. In Section 3, we present extensive numerical experiments to illustrate the empirical performance of our algorithm and compare it with other state-of-the-art methods. A real-world data analysis is presented in Section 4. Finally, we provide some concluding remarks and future directions in Section 5. Technique proofs and additional simulation results are deferred to Appendix.

2 Sparse SIR via the splicing algorithm

2.1 Notations

For any vector $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\top}\in\mathbb{R}^p$, the ℓ_q norm of $\boldsymbol{\beta}$ is defined as $\|\boldsymbol{\beta}\|_q=(\sum_{j=1}^{p}|\beta_j|^q)^{1/q}$ for $q\in\{1,2,\cdots\}$. The ℓ_0-norm of $\boldsymbol{\beta}$ is defined as $\|\boldsymbol{\beta}\|_0=\sum_{j=1}^{p}\mathbb{I}(\beta_j\neq 0)$, where $\mathbb{I}(\cdot)$ is the indicator function. Let $\mathcal{S}=\{1,\cdots,p\}$ be the full index set. For each set $\mathcal{A}\subseteq\mathcal{S}$, we denote $\mathcal{A}^c=\mathcal{S}\setminus\mathcal{A}$ as the complement of $\mathcal{A}$. Given a matrix $\boldsymbol{B}\in\mathbb{R}^{p\times q}$, we denote its ith row as $\boldsymbol{B}_{i\cdot}$ and its jth column as $\boldsymbol{B}_{\cdot j}$. We further define $\boldsymbol{B}_{\mathcal{A}\cdot}=(\boldsymbol{B}_{j\cdot},j\in\mathcal{A})$, $\boldsymbol{B}_{\cdot\mathcal{A}}=(\boldsymbol{B}_{j\cdot},j\in\mathcal{A})$, and $\boldsymbol{B}_{\mathcal{A},\mathcal{A}}=(\boldsymbol{B}_{ij})_{i\in\mathcal{A},j\in\mathcal{A}}$. For simplicity of notation, we denote $\boldsymbol{B}_{\mathcal{A}\cdot}$ as $\boldsymbol{B}_{\mathcal{A}}$, where the second subscript is omitted. We denote the support of $\boldsymbol{B}$ as $\mathrm{supp}(\boldsymbol{B})=\{j|\|\boldsymbol{B}_{j\cdot}\|_2\neq 0\}$, and we define its $\ell_{0,2}$-norm as $\|\boldsymbol{B}\|_{0,2}=\sum_{i=1}^{p}\mathbb{I}(\boldsymbol{B}_{i\cdot}=\boldsymbol{0})$ and the Frobenius norm as $\|\boldsymbol{B}\|_F=\sqrt{\sum_{i,j}\boldsymbol{B}_{ij}^2}$.

2.2 Sparse SIR as nonconvex optimization

In this subsection, let us review the setting of sparse SIR, which was formulated by Tan et al.[19] as a generalized eigenvalue decomposition problem. Without loss of generality, we first assume that the response variable Y is continuous. We then divide the range of Y into H slices $\{J_1,\cdots,J_H\}$, following the standard SIR procedure[1, 19, 28]. We define $\tilde{Y}$ as the discretized version of Y: $\tilde{Y}=\sum_{h=1}^{H}h\cdot\mathbb{I}(Y\in J_h)$. The sparse SIR is then formulated as follows:

$$\begin{aligned}&\boldsymbol{B}^{\dagger}=\underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmin}}-\mathrm{Tr}(\boldsymbol{B}^{\top}\boldsymbol{M}\boldsymbol{B}),\\&\text{s.t.}\quad\boldsymbol{B}^{\top}\boldsymbol{\Sigma}\boldsymbol{B}=I_d,\quad\text{and}\quad\|\boldsymbol{B}\|_{0,2}\leq s,\end{aligned}\tag{3}$$

where $\boldsymbol{M}=\mathrm{Cov}\{\mathbb{E}[\boldsymbol{X}|\tilde{Y}]\}$ and $\boldsymbol{\Sigma}=\mathrm{Cov}(\boldsymbol{X})$. The columns of $\boldsymbol{B}^{\dagger}$ form a basis of the central subspace. The natural sparse SIR estimator is obtained by substituting $\boldsymbol{M}$ and $\boldsymbol{\Sigma}$ with their sample versions $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{\Sigma}}$. Specifically, we define the sample design matrix as $\boldsymbol{x}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)^{\top}$, and the sample response vector as $\boldsymbol{y}=(y_1,\cdots,y_n)^{\top}$. The sample mean vector is denoted by $\bar{\boldsymbol{x}}_h$, and $\bar{\boldsymbol{x}}_h=\frac{1}{n_h}\sum_{i=1}^{n}\mathbb{I}(\tilde{y}_i=h)\boldsymbol{x}_i$ represents the sample mean in the h-th slice, where $n_h=\sum_{i=1}^{n}\mathbb{I}(\tilde{y}_i=h)$ is the number of observations in the h-th slice. The natural sparse SIR estimator is defined as follows:

$$\begin{aligned}&\boldsymbol{B}^{\star}=\underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmin}}-\mathrm{Tr}(\boldsymbol{B}^{\top}\widehat{\boldsymbol{M}}\boldsymbol{B}),\\&\text{s.t.}\quad\boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}=I_d,\quad\text{and}\quad\|\boldsymbol{B}\|_{0,2}\leq s,\end{aligned}\tag{4}$$

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

where $\widehat{\boldsymbol{M}}=\sum_{h=1}^{H}\frac{n_h}{n}(\bar{\boldsymbol{x}}_h-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_h-\bar{\boldsymbol{x}})^{\top}$ and $\widehat{\boldsymbol{\Sigma}}=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\top}$, respectively. To assess the accuracy of central subspace estimation, Tan et al.[19] introduced and examined various loss functions. In our study, we employ a general loss function defined as follows:

$$\rho(\widehat{\boldsymbol{B}},\boldsymbol{B})=\|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\top}-\boldsymbol{B}\boldsymbol{B}^{\top}\|_F^2. \tag{5}$$

As demonstrated by Ref. [19], under specific regularity conditions, the natural sparse SIR estimator attains the optimal rate with this loss function.

The upper bound of $\rho(\boldsymbol{B}^{\dagger},\boldsymbol{B}^{\star})$ is given by Theorem 2 in Ref. [19]. Before delving into the algorithm for solving the nonconvex optimization problem, we establish the following proposition, which demonstrates that the natural sparse SIR estimator successfully recovers the true support of $\boldsymbol{B}^{\dagger}$.

Proposition 1. Under regularity conditions in Ref. [19], the natural estimator successfully recovers the true support of $\boldsymbol{B}^{\star}$ if the minimum signal is large enough:

$$\mathrm{supp}(\boldsymbol{B}^{\dagger})\subseteq\mathrm{supp}(\boldsymbol{B}^{\star}),\quad \text{if}\quad \min_{j\in\mathcal{A}}\|\boldsymbol{B}^{\dagger}_{j,\cdot}\|_2^4>\rho(\boldsymbol{B}^{\dagger},\boldsymbol{B}^{\star}).$$

This proposition lays the theoretical foundation for solving the SIR problem by approximately finding $\boldsymbol{B}^{\star}$. It establishes the active set recovery property and directly implies the exact recovery, $\mathrm{supp}(\boldsymbol{B}^{\dagger})=\mathrm{supp}(\boldsymbol{B}^{\star})$, of the natural SIR estimator. Specifically, if we assume that both the the true parameter $\boldsymbol{B}^{\dagger}$ and the natural sparse SIR estimator $\boldsymbol{B}^{\star}$ have a support size of s, then $\boldsymbol{B}^{\dagger}$ exactly recovers the true support. This assumption is generally valid because increasing the support size typically leads to a decrease in the loss function. We end this section with remarks on the problem formulation and support recovery property.

Remark 1. The SIR problem has an alternative least square formulation[14]. Although more straightforward, the optimization problem of the least square formulation does not incorporate the rank-deficient structure $\mathrm{rank}(\boldsymbol{B})\le d$ inherently. This may lead to some loss of information, and such a formulation relies on n being relatively large compared to p. In contrast, Theorems 1 and 2 from Ref. [19] demonstrate that the natural sparse SIR estimator is rate optimal under three conventional loss functions, and is theoretically suitable for the "large p, small n" scenario.

Remark 2. The eigenvalue condition on the design matrix required for Proposition 1 is detailed as condition (ii) in Section 3.1 of Ref. [19]. Notably, the sample version of this condition is less stringent than the irrepresentable condition[29], which is essentially required for the SEAS method to achieve exact recovery[20].

2.3 A splicing slgorithm for sparse SIR

In this subsection, we explore the optimal conditions for problem (4) and formulate our algorithm based on these conditions. The determination of the central subspace dimension d has long been an issue. Various methods for estimating d have been proposed in the literature[1, 30, 31]. In our study, we proceed under the assumption that d is known. We initially derive our algorithm for a fixed sparsity level s and discuss the selection of s in Section 2.4. Specifically, we let $\|\boldsymbol{B}\|_{0,2}=s$, and the nonconvex optimization problem takes the following augmented form:

$$\begin{aligned}\min_{\boldsymbol{B}\in\mathbb{R}^{p\times d}}\ & -\mathrm{Tr}\left(\boldsymbol{B}^{\top}\widehat{\boldsymbol{M}}\boldsymbol{B}\right)+\frac{\rho}{2}\|\boldsymbol{C}-\boldsymbol{B}\|_F^2\\ \text{s.t.}\ \ & \boldsymbol{B}=\boldsymbol{C},\\ & \boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}=\boldsymbol{I}_d,\\ & \|\boldsymbol{C}\|_{0,2}=s,\end{aligned} \tag{6}$$

where $\rho>0$ is the regularization parameter. The Lagrangian form of the above problem is written as,

$$\begin{aligned}L_{\rho}(\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{\Lambda},\mu)=&-\mathrm{Tr}(\boldsymbol{B}^{\top}\widehat{\boldsymbol{M}}\boldsymbol{B})+\frac{\rho}{2}\|C-B\|_F^2+\langle\boldsymbol{D},\boldsymbol{C}-\boldsymbol{B}\rangle+\\ &\langle\boldsymbol{\Lambda},\boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}-\boldsymbol{I}\rangle+\mu(\|\boldsymbol{C}\|_{0,2}-s),\end{aligned} \tag{7}$$

where $\boldsymbol{D}$, $\boldsymbol{\Lambda}$, and μ are dual variables. The following proposition gives the optimal conditions for problem (6).

Proposition 2. Suppose $(\boldsymbol{B}^{\circ},\boldsymbol{C}^{\circ})$ is a row-wise minimizer of the primal optimization problem (6), and $\boldsymbol{D}^{\circ},\boldsymbol{\Lambda}^{\circ}$, and μ° are associated dual variables. Denote $\mathcal{A}^{\circ}=\{j|\|\boldsymbol{B}^{\circ}_{j,\cdot}\|_2\neq 0\}$ and $\mathcal{I}^{\circ}=(\mathcal{A}^{\circ})^{c}$. We refer to $\mathcal{A}$ and $\mathcal{I}$ as the active and inactive sets, respectively. Then $(\boldsymbol{B}^{\circ},\boldsymbol{C}^{\circ},\boldsymbol{D}^{\circ},\boldsymbol{\Lambda}^{\circ},\mu^{\circ})$ and $(\mathcal{A}^{\circ},\mathcal{I}^{\circ})$ satisfy the following conditions:

$$\begin{aligned}&\boldsymbol{B}^{\circ}_{\mathcal{A}^{\circ}}=\underset{\boldsymbol{B}\in\mathbb{R}^{s\times d}}{\mathrm{argmin}}\left\{-\mathrm{Tr}(\boldsymbol{B}^{\top}\widehat{\boldsymbol{M}}_{\mathcal{A}^{\circ},\mathcal{A}^{\circ}}\boldsymbol{B}):\boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^{\circ},\mathcal{A}^{\circ}}\boldsymbol{B}=I_d\right\},\quad \boldsymbol{B}^{\circ}_{\mathcal{I}^{\circ}}=\boldsymbol{0},\\ &\boldsymbol{D}^{\circ}_{\mathcal{A}^{\circ}}=\boldsymbol{0},\quad \boldsymbol{D}^{\circ}_{\mathcal{I}^{\circ}}=-2\widehat{\boldsymbol{M}}\boldsymbol{B}^{\circ}+2\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}^{\circ}\boldsymbol{\Lambda}^{\circ},\\ &\Lambda^{\circ}=\mathrm{diag}((\boldsymbol{B}^{\circ}_{\cdot,1})^{\top}\widehat{\boldsymbol{M}}\boldsymbol{B}^{\circ}_{\cdot,1},\cdots,(\boldsymbol{B}^{\circ}_{\cdot,d})^{\top}\widehat{\boldsymbol{M}}\boldsymbol{B}^{\circ}_{\cdot,d}),\\ &\boldsymbol{C}^{\circ}=\boldsymbol{B}^{\circ},\\ &\mathcal{A}^{\circ}=\left\{i|\sum_{k}\mathbb{I}\left(\|\boldsymbol{B}^{\circ}_{i,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^{\circ}_{i,\cdot}\|_2\le\|\boldsymbol{B}^{\circ}_{k,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^{\circ}_{k,\cdot}\|_2\right)\le s\right\}.\end{aligned} \tag{8}$$

Remark 3. Due to the nonconvex nature of the optimization problem, it may have multiple local minimizers. In such circumstances, the literature on cardinality-constrained algorithms usually considers coordinate-wise or row-wise minimizers. For more details, we refer to Refs. [32–34].

Proposition 2 establishes the groundwork for developing our algorithm, which employs an iterative approach to approximate the optimal conditions. The proof is given later in Section 2.5.

Let $\{\boldsymbol{B}^m,\boldsymbol{D}^m,\boldsymbol{\Lambda}^m\}$ denote the corresponding values of the variables at the mth iteration. We first update the active set by

$$\mathcal{A}^{m+1}=\left\{i|\sum_{k}\mathbb{I}\left(\|\boldsymbol{B}^{m}_{i,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^{m}_{i,\cdot}\|_2\le\|\boldsymbol{B}^{m}_{k,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^{m}_{k,\cdot}\|_2\right)\le s\right\},$$

and let $\mathcal{I}^{m+1}=(\mathcal{A}^{m+1})^{c}$. Then we update the primal variable $\boldsymbol{B}$ and dual variable $\boldsymbol{D}$. Specifically, we update $\boldsymbol{D}^{m+1}_{\mathcal{A}^{m+1}}=\boldsymbol{0}$, $\boldsymbol{B}^{m+1}_{\mathcal{I}^{m+1}}=\boldsymbol{0}$, and

$$\begin{aligned}&\boldsymbol{B}^{m+1}_{\mathcal{A}^{m+1}}=\underset{\boldsymbol{B}\in\mathbb{R}^{s\times d}}{\mathrm{argmin}}-\mathrm{Tr}(\boldsymbol{B}^{\top}\widehat{\boldsymbol{M}}_{\mathcal{A}^{m+1},\mathcal{A}^{m+1}}\boldsymbol{B}),\\ &\text{s.t.}\quad \boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^{m+1},\mathcal{A}^{m+1}}\boldsymbol{B}=I_d.\end{aligned}$$

This is a generalized eigenvalue decomposition problem. We can also obtain the generalized eigenvalues

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

$$\Lambda^{m+1} = \mathrm{diag}((\boldsymbol{B}^{m+1}_{\cdot,1})^\top \widehat{\boldsymbol{M}} \boldsymbol{B}^{m+1}_{\cdot,1}, \cdots, (\boldsymbol{B}^{m+1}_{\cdot,d})^\top \widehat{\boldsymbol{M}} \boldsymbol{B}^{m+1}_{\cdot,d}).$$

Then we update $\boldsymbol{D}^{m+1}_{\mathcal{I}^{m+1}} = -2\widehat{\boldsymbol{M}}\boldsymbol{B}^{m+1} + 2\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}^{m+1}\boldsymbol{\Lambda}^{m+1}$.

The regularization parameter ρ tunes the "step size" in each iteration. Intuitively, a large ρ leads to slow updates of the active set, while a small ρ leads to fast updates. The update procedure of the active set can be interpreted as an exchange between the active and inactive sets. This allows us to design the algorithm in a splicing manner, wherein the "step size" is adaptively determined based on the exchange size, denoted by T. It is evident that a specific ρ can determine a splicing size T. However, the challenge lies in identifying the range of ρ for a predetermined T. Let $S^{m+1}_{T,1}$ and $S^{m+1}_{T,2}$ represent the exchange sets for the active and inactive sets, respectively. We obtain,

$$\begin{aligned}
S^{m+1}_{T,1} =& \{i \in \mathcal{A}^m | \sum_{k\in\mathcal{A}^m} \mathbb{I}(\|\boldsymbol{B}^m_{i,\cdot} - \frac{1}{\rho}\boldsymbol{D}^m_{i,\cdot}\|_2 \geq \mathbb{I}(\|\boldsymbol{B}^m_{k,\cdot} - \frac{1}{\rho}\boldsymbol{D}^m_{k,\cdot}\|_2) \leq T\} = \\
& \{i \in \mathcal{A}^m | \sum_{k\in\mathcal{A}^m} \mathbb{I}(\|\boldsymbol{B}^m_{i,\cdot}\|_2 \geq \mathbb{I}(\|\boldsymbol{B}^m_{k,\cdot}\|_2) \leq T\}, \\
S^{m+1}_{T,2} =& \{i \in \mathcal{I}^m | \sum_{k\in\mathcal{I}^m} \mathbb{I}(\|\boldsymbol{B}^m_{i,\cdot} - \frac{1}{\rho}\boldsymbol{D}^m_{i,\cdot}\|_2 \geq \mathbb{I}(\|\boldsymbol{B}^m_{k,\cdot} - \frac{1}{\rho}\boldsymbol{D}^m_{k,\cdot}\|_2) \leq T\} = \\
& \{i \in \mathcal{I}^m | \sum_{k\in\mathcal{I}^m} \mathbb{I}(\|\boldsymbol{D}^m_{i,\cdot}\|_2 \geq \mathbb{I}(\|\boldsymbol{D}^m_{k,\cdot}\|_2) \leq T\}.
\end{aligned} \tag{9}$$

Since the active set is the indices of the largest s row norms of $\boldsymbol{B}^m - \frac{1}{\rho}\boldsymbol{D}^m$, the exchange set in the active set $S^{m+1}_{T,1}$ is the indices of the smallest T row norms of $\boldsymbol{B}^m - \frac{1}{\rho}\boldsymbol{D}^m$, and the exchange set in the inactive set $S^{m+1}_{T,2}$ is the indices of the largest T row norms of $\boldsymbol{B}^m - \frac{1}{\rho}\boldsymbol{D}^m$. Moreover, the ranks of the row norms of $\boldsymbol{B}^m - \frac{1}{\rho}\boldsymbol{D}^m$ depend solely on $\boldsymbol{B}^m$ in the active set and $\boldsymbol{D}^m$ in the inactive set. Intuitively, for variable i in the active set, $\|\boldsymbol{B}^m_{i,\cdot}\|_2$ serves as the relevance of this variable, and for variable j in the inactive set, $\|\boldsymbol{D}^m_{i,\cdot}\|_2$ serves as the relevance of this variable. The procedure of our algorithm involves iteratively exchanging the least relevant elements currently in the active set with the most pertinent elements from the inactive set.

The magnitude of ρ influences only the size of the exchange sets. The following proposition establishes the equivalence between determining the regularization parameter ρ and determining the exchange size T.

Proposition 3. Given exchange size T, the range of ρ is given by

$$\rho \in \begin{cases} \left(\dfrac{\min_{i\in S^{m+1}_{T,2}}\|\boldsymbol{D}^m_{i,\cdot}\|}{\max_{i\in S^{m+1}_{T,1}}\|\boldsymbol{B}^m_{i,\cdot}\|}, +\infty\right), & T = 0; \\ \left(\dfrac{\min_{i\in S^{m+1}_{T+1,2}}\|\boldsymbol{D}^m_{i,\cdot}\|}{\max_{i\in S^{m+1}_{T,1}}\|\boldsymbol{B}^m_{i,\cdot}\|}, \dfrac{\min_{i\in S^{m+1}_{T+1,2}}\|\boldsymbol{D}^m_{i,\cdot}\|}{\max_{i\in S^{m+1}_{T,1}}\|\boldsymbol{B}^m_{i,\cdot}\|}\right], & 1 \leq T < s; \\ \left(0, \dfrac{\min_{i\in S^{m+1}_{T,2}}\|\boldsymbol{D}^m_{i,\cdot}\|}{\max_{i\in S^{m+1}_{T,1}}\|\boldsymbol{B}^m_{i,\cdot}\|}\right], & T = s. \end{cases}$$

Proposition 3 implies that we can convert the problem of tuning the regularization parameter ρ into determining the optimal exchange size T, a much easier task. To be specific, we adopt a natural approach to determine T in a data-driven manner, that is, we choose $T \in \{1, 2, \cdots, T_{\max}\}$ such that the objective function can maximally increase, where $T_{\max}$ is a fixed maximum exchange size. Finally, our algorithm terminates when no more increase in objective function can be achieved. The entire algorithmic procedure is summarized as Algorithm 1.

2.4 Implementation details

We have implemented the proposed algorithm in R, which is freely available at https://github.com/brtang63/A-Splicing-Algorithm-for-Best-Subset-Selection-in-Sliced-Inverse-Regression. While the R package **abess**[35] offers an efficient implementation of splicing techniques across a broad range of scenarios, our implementation adds a new scenario specifically tailored to address unique challenges in SIR. In the following, we describe details of our implementation.

Initialization. In Algorithm 1, we need to specify an initial active set $\mathcal{A}_0$. As discussed above, given the parameter $\boldsymbol{B}$, $\|\boldsymbol{B}_{i,\cdot}\|_2$ is the relevance of variable i in the active set. This in-

Algorithm 1. Splicing-SIR.

Require: $\widehat{\boldsymbol{M}}$, $\widehat{\boldsymbol{\Sigma}}$, the dimension of the central subspace d, the sparsity level s, the initial active set $\mathcal{A}^0$, the maximum splicing size $T_{\max}$, and the maximum iteration number $m_{\max}$.

1: Initialize:
$\mathcal{I}^0 = (\mathcal{A}^0)^c$, $\boldsymbol{B}^0_{\mathcal{A}^0} = \underset{\boldsymbol{B}\in\mathbb{R}^{s\times d}}{\mathrm{argmin}}\left\{-\mathrm{Tr}(\boldsymbol{B}^\top \widehat{\boldsymbol{M}}_{\mathcal{A}^0,\mathcal{A}^0}\boldsymbol{B}) : \boldsymbol{B}^\top\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}^0,\mathcal{A}^0}\boldsymbol{B} = I_d\right\}$,
$\boldsymbol{B}^0_{\mathcal{I}^0} = \boldsymbol{0}$, $\Lambda^0 = \mathrm{diag}((\boldsymbol{B}^0_{\cdot,1})^\top \widehat{\boldsymbol{M}}\boldsymbol{B}^0_{\cdot,1}, \cdots, (\boldsymbol{B}^0_{\cdot,d})^\top \widehat{\boldsymbol{M}}\boldsymbol{B}^0_{\cdot,d})$,
$\boldsymbol{D}_{\mathcal{I}^0} = (-2\widehat{\boldsymbol{M}}\boldsymbol{B}^0 + 2\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}^0\boldsymbol{\Lambda}^0)_{\mathcal{I}^0}$, $\boldsymbol{D}^0_{\mathcal{A}^0} = \boldsymbol{0}$.
2: **for** $m = 0, \cdots, m_{\max}$ **do**
3: $(\mathcal{A}^{m+1}, \mathcal{I}^{m+1}, \boldsymbol{B}^{m+1}, \boldsymbol{D}^{m+1}, \boldsymbol{\Lambda}^{m+1}) = (\mathcal{A}^m, \mathcal{I}^m, \boldsymbol{B}^m, \boldsymbol{D}^m, \boldsymbol{\Lambda}^m)$.
4: Compute the objective function: $L = \mathrm{Tr}(\boldsymbol{B}^{m\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}^m)$.
5: **for** $T = 1, \cdots, T_{\max}$ **do**
6: Compute candidate exchange sets $S^{m+1}_{T,1}$ and $S^{m+1}_{T,2}$ by (9), and update candidate active and inactive sets: $\widetilde{\mathcal{A}} = (\mathcal{A}^m \backslash S^{m+1}_{T,1}) \cup S^{m+1}_{T,2}$, $\widetilde{\mathcal{I}} = (\widetilde{\mathcal{A}})^c$.
7: Compute candidate parameters:
$\widetilde{\boldsymbol{B}}_{\widetilde{\mathcal{A}}} = \underset{\widetilde{\boldsymbol{B}}\in\mathbb{R}^{s\times d}}{\mathrm{argmin}}\left\{-\mathrm{Tr}(\widetilde{\boldsymbol{B}}^\top \widehat{\boldsymbol{M}}_{\widetilde{\mathcal{A}},\widetilde{\mathcal{A}}}\widetilde{\boldsymbol{B}}) : \widetilde{\boldsymbol{B}}^\top\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{A}},\widetilde{\mathcal{A}}}\widetilde{\boldsymbol{B}} = I_d\right\}$, $\widetilde{\boldsymbol{B}}_{\widetilde{\mathcal{I}}} = \boldsymbol{0}$.
$\widetilde{\Lambda} = \mathrm{diag}((\widetilde{\boldsymbol{B}}_{\cdot,1})^\top \widehat{\boldsymbol{M}}\widetilde{\boldsymbol{B}}_{\cdot,1}, \cdots, (\widetilde{\boldsymbol{B}}_{\cdot,d})^\top \widehat{\boldsymbol{M}}\widetilde{\boldsymbol{B}}_{\cdot,d})$,
$\widetilde{\boldsymbol{D}}_{\widetilde{\mathcal{I}}} = (-2\widehat{\boldsymbol{M}}\widetilde{\boldsymbol{B}} + 2\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{\Lambda}})_{\widetilde{\mathcal{I}}}$, $\widetilde{\boldsymbol{D}}_{\widetilde{\mathcal{A}}} = \boldsymbol{0}$.
8: Compute the candidate objective function: $\widetilde{L} = \mathrm{Tr}(\widetilde{\boldsymbol{B}}^\top\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{B}})$.
9: **if** $\widetilde{L} > L$ **then**
10: $L = \widetilde{L}, (\mathcal{A}^{m+1}, \mathcal{I}^{m+1}, \boldsymbol{B}^{m+1}, \boldsymbol{D}^{m+1}, \boldsymbol{\Lambda}^{m+1}) = (\widetilde{\mathcal{A}}, \widetilde{\mathcal{I}}, \widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{D}}, \widetilde{\boldsymbol{\Lambda}})$.
11: **end if**
12: **end for**
13: **if** $(\mathcal{A}^{m+1}, \mathcal{I}^{m+1}) = (\mathcal{A}^m, \mathcal{I}^m)$ **then**
14: Break from the **for** loop.
15: **end if**
16: **end for**

Ensure: $(\mathcal{A}^{m+1}, \boldsymbol{B}^{m+1})$

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

tuition leads to the following naive method for initialization. We apply screening for the top-s rows of B, where B is an initial non-sparse estimator obtained through generalized eigenvalue decomposition. To be specific,

(i) Apply generalized eigenvalue decomposition to $(\widehat{\boldsymbol{M}},\widehat{\boldsymbol{\Sigma}})$, obtain the top d generalized eigenvectors $\boldsymbol{B}=(\beta_1,\cdots,\beta_d)$.

(ii) Calculate the magnitude of each row $\|\boldsymbol{B}_{i,\cdot}\|_2$ for $i\in\{1,2,\cdots,p\}$.

(iii) Set the indices of the largest s values as the initial active set.

Parameter tuning. To determine the sparsity level s, we apply K-fold cross-validation. We use the distance correlation[36], denoted as $\mathrm{dCor}(Y,\boldsymbol{B}^\top\boldsymbol{X})$, as the cross-validation loss function, which is widely used in the SDR problem[20]. The validity of this evaluation measure was inspired by Sheng and Yin[37], who established that under the normality constraint and mild conditions, the distance covariance between Y and $\boldsymbol{B}^\top\boldsymbol{X}$ is maximized at the basis of the central subspace.

The number of slices H is another hyperparameter to be determined. Li[1] and Wu et al[14] demonstrated that inverse mean-based methods are not excessively sensitive to the choice of H, and we set $H=5$ in our numerical studies, which is also a typical setting in the literature[14].

2.5 Derivation of optimal conditions

Proof of Proposition 2. Recall the Lagrangian function $L_\rho(\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{\Lambda},\mu)$ defined in (7). Deriving the stationary condition for parameter $\boldsymbol{B}$ is straightforward. Let $\frac{\partial L_\rho}{\partial \boldsymbol{B}^\circ}=\boldsymbol{0}$, we obtain the necessary condition

$$2\boldsymbol{\Sigma}\boldsymbol{B}^\circ\boldsymbol{\Lambda}^\circ-2\boldsymbol{M}\boldsymbol{B}^\circ-\rho(\boldsymbol{C}^\circ-\boldsymbol{B}^\circ)-\boldsymbol{D}^\circ=\boldsymbol{0}. \tag{10}$$

However, L_ρ is non-differentiable with respect to parameter $\boldsymbol{C}$ due to the carnality constraint. To obtain the stationary condition, define

$$F_\rho(\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\mu)=\frac{\rho}{2}\|\boldsymbol{C}-\boldsymbol{B}\|_F^2+\langle\boldsymbol{D},\boldsymbol{C}-\boldsymbol{B}\rangle+\mu\|\boldsymbol{C}\|_{0,2},$$

which is obtained by removing items in L_ρ irrelevant to C. The row-wise minimizer is given in the following lemma.

Lemma 1. If $\boldsymbol{C}^\circ$ is the row-wise minimizer of function F_ρ, then it satisfies

$$\boldsymbol{C}^\circ=H_{\frac{2\mu^\circ}{\rho}}(\boldsymbol{B}^\circ-\frac{1}{\rho}\boldsymbol{D}^\circ), \tag{11}$$

where $H_t(\cdot)$ denotes the row-wise hard thresholding operator defined as,

$$(H_t(\cdot))_{i,\cdot}=\begin{cases}0, & \text{if}\quad \|\boldsymbol{X}_{i,\cdot}\|_2<\sqrt{t};\\ \boldsymbol{X}_{i,\cdot}, & \text{if}\quad \|\boldsymbol{X}_{i,\cdot}\|_2\geq\sqrt{t}.\end{cases}$$

Proof of Lemma 1 is deferred to Appendix A.1. Combining Eq. (11) in Lemma 1, Eq. (10), and the equality constraints in problem (6) we obtain the necessary conditions as follows:

$$\begin{aligned}&2\boldsymbol{\Sigma}\boldsymbol{B}^\circ\boldsymbol{\Lambda}^\circ-2\boldsymbol{M}\boldsymbol{B}^\circ-\rho(\boldsymbol{C}^\circ-\boldsymbol{B}^\circ)-\boldsymbol{D}^\circ=\boldsymbol{0},\\ &\boldsymbol{C}^\circ=H_{\frac{2\mu^\circ}{\rho}}(\boldsymbol{B}^\circ-\frac{1}{\rho}\boldsymbol{D}^\circ),\\ &\boldsymbol{C}^\circ=\boldsymbol{B}^\circ,\quad \boldsymbol{B}^{\circ\top}\boldsymbol{\Sigma}\boldsymbol{B}^\circ=\boldsymbol{I},\quad \|\boldsymbol{C}^\circ\|_{0,2}=s.\end{aligned} \tag{12}$$

Accordingly, we have

$$\begin{aligned}&\boldsymbol{D}^\circ=-2\boldsymbol{M}\boldsymbol{B}^\circ+2\boldsymbol{\Sigma}\boldsymbol{B}^\circ\boldsymbol{\Lambda}^\circ,\\ &\boldsymbol{B}^\circ=H_{\frac{2\mu^\circ}{\rho}}(\boldsymbol{B}^\circ-\frac{1}{\rho}\boldsymbol{D}^\circ),\\ &\boldsymbol{B}^{\circ\top}\boldsymbol{\Sigma}\boldsymbol{B}^\circ=\boldsymbol{I},\quad \|\boldsymbol{B}^\circ\|_{0,2}=s.\end{aligned} \tag{13}$$

Moreover, by definition of $\mathcal{A}^\circ$ and the second equality of (13), we have

$$\mathcal{A}^\circ=\left\{i\middle|\sum_k\mathbb{I}\left(\|\boldsymbol{B}^\circ_{i,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^\circ_{i,\cdot}\|_2\leq\|\boldsymbol{B}^\circ_{k,\cdot}-\tfrac{1}{\rho}\boldsymbol{D}^\circ_{k,\cdot}\|_2\right)\leq s\right\}.$$

Now the optimal conditions are straightforward: The first line of (8) follows from problem (4) and the definition of $\mathcal{A}$, and the second line of (8) is directly derived from the first equation of (13). Finally, note that

$$\boldsymbol{D}^\circ_{\mathcal{A}}=(-2\boldsymbol{M}\boldsymbol{B}^\circ+2\boldsymbol{\Sigma}\boldsymbol{B}^\circ\boldsymbol{\Lambda}^\circ)_{\mathcal{A}}\quad\text{and}\quad\boldsymbol{D}^\circ_{\mathcal{A}}=\boldsymbol{0},$$

the third line of (8) follows immediately.

3 Simulation

This section elaborates on the numerical experiments carried out to assess our algorithm's empirical performance. In particular, we benchmark the performance of our method with that of other state-of-the-art methods by employing multiple evaluation criteria. Through such comparisons, we comprehensively examine both the computational efficiency and estimation accuracy of our algorithm.

3.1 Simulation settings

We consider the following models.

(a) $Y=\dfrac{\boldsymbol{X}^\top\boldsymbol{\beta}_1}{0.5+(1.5+\boldsymbol{X}^\top\boldsymbol{\beta}_2)^2}+0.2\epsilon.$

(b) $Y=(\boldsymbol{X}^\top\boldsymbol{\beta}_1)\cdot\exp(\boldsymbol{X}^\top\boldsymbol{\beta}_2+0.5\epsilon).$

Here ϵ is a noise variable following the normal distribution $\mathcal{N}(0,1)$. These two models with two-dimensional central subspaces were previously considered in the numerical experiments of Ref. [14] and Ref. [20], separately. The simulation results for models with one-dimensional central subspaces are provided in Appendix. We compare our splicing-SIR method with existing methods, including LASSO-SIR[15], SEAS-SIR[20], and Relaxed Natural SSIR — the natural sparse SIR estimator solved via a relaxed optimization problem, as studied by Tan et al. [19]

Data Generation. Each row of the design matrix is independently sampled from a normal distribution $N(\boldsymbol{0},\boldsymbol{\Sigma})$, with two types of covariance structures considered: an independent structure ($\boldsymbol{\Sigma}=\boldsymbol{I}$) and a correlated structure ($\boldsymbol{\Sigma}_{ij}=0.5^{\mathbb{I}(i\neq j)}$). For the coefficients $\boldsymbol{\beta}$, the first four elements of the first column and elements five to eight of the second column are assigned a value of 0.5: specifically, $\boldsymbol{\beta}_{1:4,1}=0.5$ and $\boldsymbol{\beta}_{5:8,2}=0.5$. All other elements $\boldsymbol{\beta}_{i,j}$ are set to 0.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

Evaluation Criteria. To evaluate the ability of various methods to uncover the true model and the computational efficiency, we employ the following criteria.

① **Subset recovery probability.** To quantify the accuracy of subset selection, we measure the frequency with which the method correctly recovers the active set, the inactive set, and both sets. The performance metrics for these criteria are mathematically expressed as follows: the active set recovery probability $\mathbb{P}(\mathcal{A}^\star \subseteq \widehat{\mathcal{A}})$, the inactive set recovery probability $\mathbb{P}(\mathcal{I}^\star \subseteq \widehat{\mathcal{I}})$ and the exact subset recovery probability $\mathbb{P}(\mathcal{A}^\star = \widehat{\mathcal{A}})$.

② **Sparsity level.** We investigate the estimated sparsity levels of the different methods. Additionally, according to our data generation settings, the true sparsity level is 8.

③ **Vector correlation coefficient.** To evaluate the accuracy of subspace estimation, we employ the canonical correlation coefficient, as introduced in Ref. [38]. It was later used by Wu and Li [14] to assess subspace estimation performance.

This coefficient measures the correlation between the basis of the estimated and true central subspaces, providing a formal metric for comparison. It is defined as follows:

$$\rho = \max_{\boldsymbol{u},\boldsymbol{v}} \mathrm{Cor}(\boldsymbol{u}^\top \widehat{\boldsymbol{B}}, \boldsymbol{v}^\top \boldsymbol{B}^\star),$$

where $\mathrm{Cor}(\cdot,\cdot)$ denotes Pearson's correlation.

④ **Runtime.** We compare the runtimes of all the methods involved. In particular, all the experiments were carried out on a Linux platform with AMD EPYC 9654 96-Core Processors.

3.2 Statistical performance

We analyze our algorithm across various sample sizes, focusing on a fixed dimension of $p = 600$. The sample size is progressively increased from 1000 to 3000, in increments of 500. The simulation results for the high-dimensional scenarios are provided in Appendix. Our investigation is based on 200 randomly generated datasets. The statistical performance under Model (a) and Model (b), with the dimension set at $p = 600$, is shown in Figs. 1 and 2, respectively.

In the context of an independent design setting, splicing-SIR and Relaxed Natural SSIR exhibit competitive performances and substantially outperform SEAS and LASSO-SIR. Specifically, all methods identify the best subset in most cases. Splicing-SIR and Relaxed Natural SSIR demonstrate a higher success rate, exceeding 0.8 for sample sizes above 1500, in precisely identifying the best subset. In contrast, SEAS-SIR achieves this with a probability below 0.5, while LASSO-SIR invariably incorporates irrelevant variables in every replicate.

The performance of different methods can be further evaluated through sparsity level considerations. Splicing-SIR and Relaxed Natural SSIR typically yield the correct sparsity level. In contrast, SEAS often results in larger sparsity levels, while LASSO-SIR always produces substantially larger models. However, as indicated by the canonical correlations, larger models do not necessarily lead to improved performance. Notably, splicing-SIR and Relaxed Natural SSIR, which are more effective in identifying the best subset, also achieve superior subspace estimation.

Splicing-SIR and Relaxed Natural SSIR show comparable performance in the independent design setting. However, when dealing with the correlated design setting, splicing-SIR continues to exhibit strong performance in both subspace and parameter estimation. In the meantime, Relaxed Natural SSIR sometimes includes irrelevant variables. In all replicates, Relaxed Natural SSIR results in a sparsity level larger than the ground truth, including some irrelevant variables. Consequently, its canonical correlation is lower than that of the splicing-SIR method.

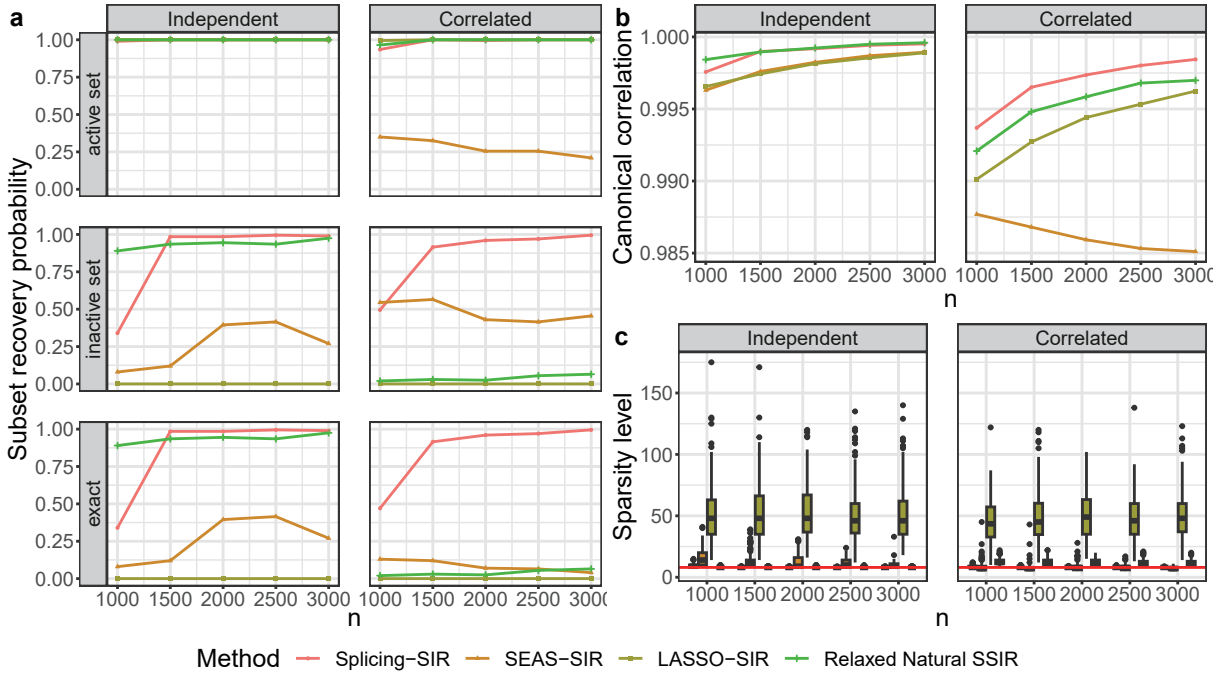


Fig. 1. Statistical performance under Model (a). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

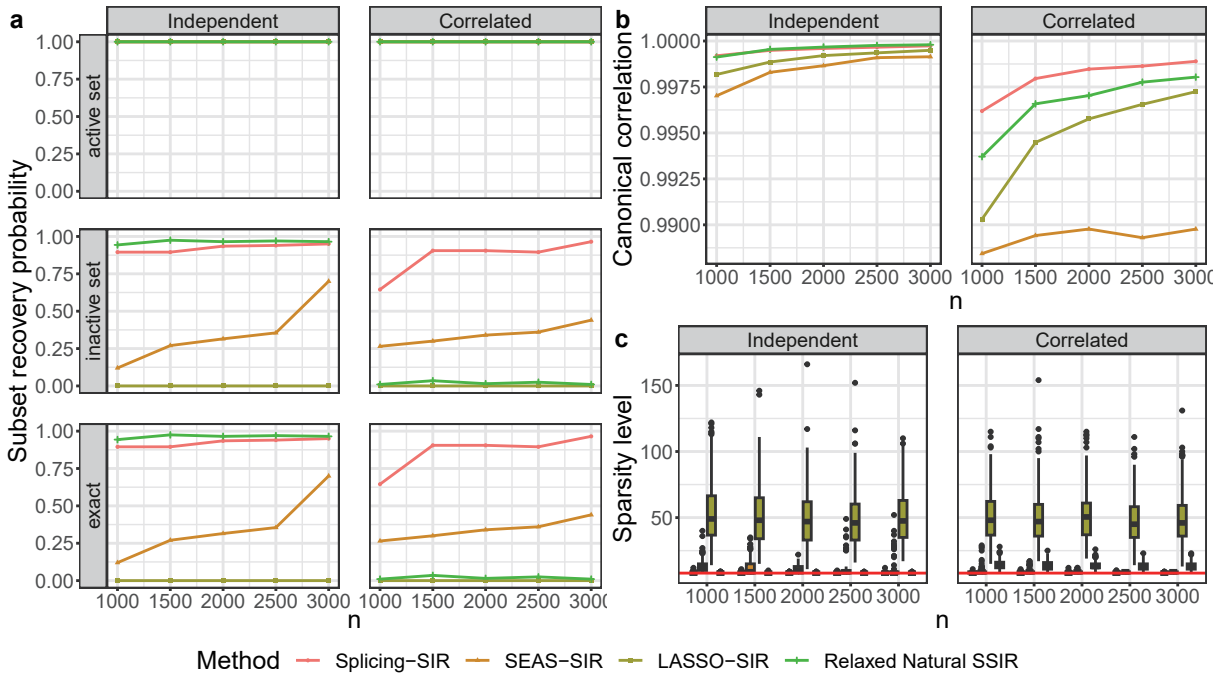


Fig. 2. Statistical performance under Model (b). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

The performance of splicing-SIR —precisely identifying the best subset, accurately estimating the central subspace, and demonstrating resilience to high correlations—are expected outcomes. This empirical achievement of splicing iterations has been validated in earlier studies, such as those by Zhu et al.[24] for linear models and Tang et al.[27] for single-index models. Additionally, Guo et al.[23] illustrated that best subset selection inherently possesses robustness against design dependencies.

3.3 Computation time

In this subsection, we compare the computation time across different methods. We set the dimension $p = 200$ and we increase the sample size n from 400 to 2000. Additionally, with a fixed sample size of $n = 1000$, we increase the dimension p from 200 to 1000. The median runtime under 100 replicates is presented in Fig. 3. Splicing-SIR is comparable to the fast LASSO-SIR method in terms of runtime and is faster than SEAS-SIR and Relaxed Natural SSIR. In particular, the runtime of SSIR increases rapidly as the dimension increases. The runtime of splicing-SIR, in contrast, is approximately linear with respect to n or p. Thus, Splicing-SIR is scalable for high-dimensional and large-scale data.

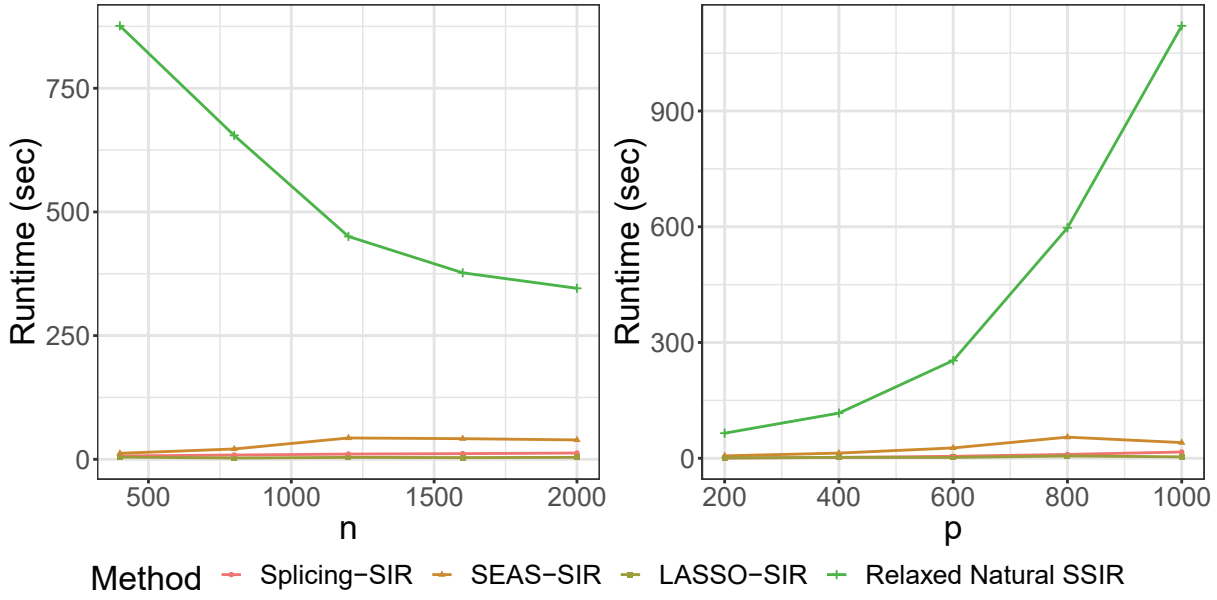


Fig. 3. (a) Runtime (y-axis) versus sample size plot. (b) Runtime (y-axis) versus dimension plot.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

4 Real data analysis

In this section, we analyze a lymphoma dataset. Specifically, we illustrate the practical application of splicing-SIR and highlight its advantages over competing methods. This dataset was previously studied in Refs. [20, 39], and we accessed it from the R package **spls** at https://CRAN.R-project.org/package=spls. This dataset included 62 samples across three lymphoma categories: 42 diffuse large B-cell lymphoma (DLBCL) samples, 9 follicular lymphoma (FL) samples, and 11 chronic lymphocytic leukemia (CLL) samples. The lymphoma type served as the response variable, and was coded as 0, 1, or 2 for DLBCL, FL, or CLL, respectively. Each sample was characterized by 4026 gene expression measurements.

Given the ultra-high dimensionality of the dataset, we first screened the predictors by distance correlation and retained the 100 most relevant variables. Since the responses were categorical, we calculated the distance correlations with one-hot response coding following common practices[20]. Fig. 4 illustrates the pairwise correlation structure of these 100 variables. Most variable pairs exhibit correlation coefficients exceeding 0.5, indicating high correlations among the predictors. We applied the aforementioned methods to analyze this dataset. According to the suggestion of Zeng et al.[20], the central subspace is two-dimensional. We configure all methods with the setting of $d = 2$.

Each method leads to an estimated central subspace. To intuitively present the dimension reduction results, we visualize the reduced-dimensional predictors on the central subspace, a two-dimensional plane consisting of two linear predictors generated by linear combinations of a subset of predictors. The results are shown in Fig. 5. To evaluate the dimension reduction performance, we fit a multinomial logistic regression for the response using these two linear predictors and record the separability of the three classes in Table 1. The sparsity levels and runtime in seconds are also recorded in it.

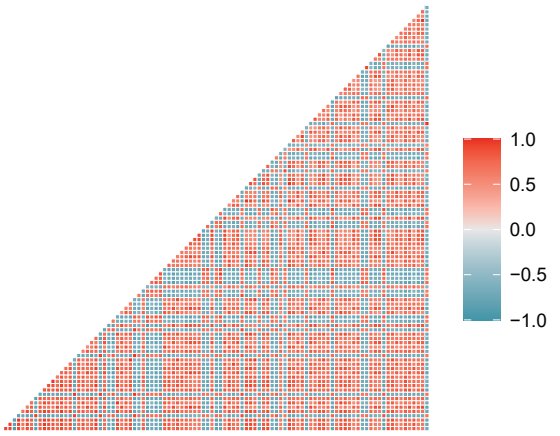


Fig. 4. Pairwise correlation structure of the design matrix.

As displayed in Table 1, all methods, with the exception of Relaxed Natural SSIR, achieve accurate classification. Notably, splicing-SIR attains perfect classification using only 10 predictors. Fig. 5 visually demonstrates the central subspace of each method by showing the projected scatter points and the decision boundaries for classification. In the case of splicing-SIR, the samples from the three different types are distinctly separated by radial lines. Although SEAS–SIR and LASSO–SIR also manage to achieve perfect classification, they involve a substantially greater number of variables, po-

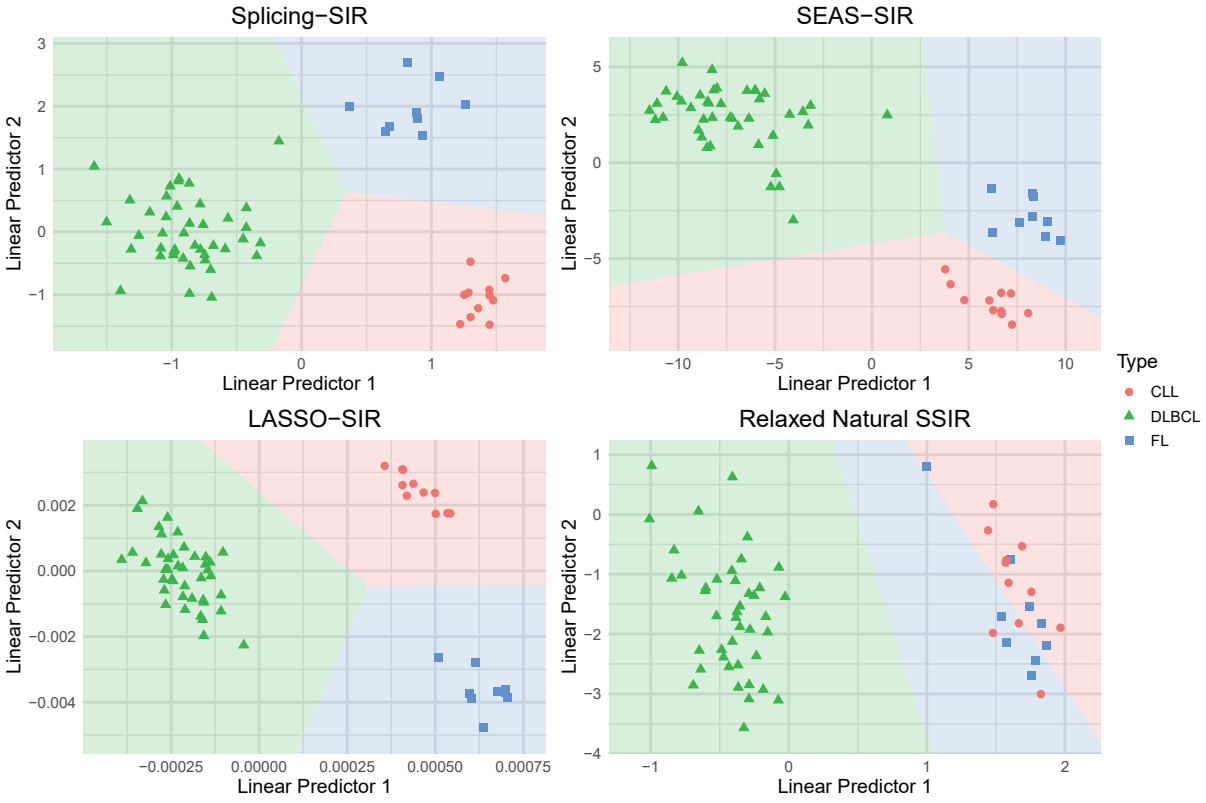


Fig. 5. The scatter plots and multinomial logistic decision boundaries on the central subspace.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

Table 1. Results of different methods on the lymphoma dataset.

Method	Sparsity level	Separable	Runtime (s)
Splicing-SIR	10	✓	1.28
SEAS-SIR	30	✓	2.55
LASSO-SIR	30	✓	0.43
Relaxed Natural SSIR	60	✗	25.2

tentially leading to less parsimonious models. This issue is probably due to the high correlations among predictors. Our simulations indicate that SEAS–SIR and LASSO–SIR could be affected by design dependence, as they might include irrelevant variables in such cases. In contrast, splicing-SIR shows robustness against high correlations. Furthermore, Relaxed Natural SSIR exhibits limitations on this dataset: it selects the densest model, takes the most time, and results in the least effective central subspace—where the three types of samples cannot be distinguished using multinomial logistic regression.

5 Conclusions

In this paper, we propose a splicing-type algorithm for best subset selection in SIR. Our method is distinguished from existing SIR methods because we directly tackle two nonconvex constraints: the sparsity constraint and the orthogonal constraint. By iteratively replacing relevant variables with irrelevant variables, our algorithmic solutions effectively approximate the optimal conditions. The empirical success of our algorithm has been shown in our numerical studies. First, while we solve the natural sparse SIR estimator proposed in Ref. [19], our algorithm is more computationally efficient. Second, our algorithm exhibits accuracy in exact support recovery, high performance in central subspace estimation, and robustness against correlations among predictors. While relaxation methods may lead to less optimal solutions[23], splicing-type algorithms achieve accurate support recovery even when predictors are correlated.

Notably, the splicing-SIR algorithm distinguishes it from the existing best subset selection literature on $\ell_{0,2}$-constrained problems in some aspects. First, unlike the scenarios in best subset of groups selection[33] and integrative analysis[40], our problem formulation incorporates an orthogonal constraint alongside the sparsity constraint, and lacks a closed-form solution even with a given active set. The development of the row-wise optimal conditions for the SIR problem, as detailed in Proposition 2, is also innovative. Second, while the primal-dual active selection (PDAS) algorithm for reduced rank regression developed by Wen et al.[25] shares similarities with the splicing method, our splicing algorithm essentially generalizes the PDAS algorithm. Specifically, the splicing algorithm introduces an additional weight ρ into the "measure of importance" in PDAS, i.e., $\|\boldsymbol{B}^m_{i,\cdot}\|_2+\rho^{-1}\|\boldsymbol{D}^m_{i,\cdot}\|_2$. Unlike the fixed setting of $\rho=1$ in PDAS, we determine the optimal ρ via the exchange size T in a data-driven manner.

There are substantial issues to be addressed in the future. While our algorithm has promising empirical performance, some theoretical aspects are not well understood. The justification of the theoretical SIR estimator has been demonstrated in Proposition 1 and Ref. [19]. However, the theoretical properties of our algorithmic solution remain unclear. Recently, theoretical aspects of the algorithmic solution for other splicing-type algorithms have been studied in papers including Refs. [24, 25]. This motivates us to further explore the theory of our algorithm in our future research. Moreover, under the linear model, Ref. [23] theoretically characterized the robustness against design dependence for both the theoretical best subset selection estimator and an approximate estimator[41]. In particular, Ref. [23] established selection consistency for the best subset selection estimator without restricted eigenvalue conditions on the design matrix. Given the theoretical results in Ref. [23] and the empirical results in this paper, theoretically characterizing the robustness against design dependence of the splicing-SIR method is a valuable future direction.

Acknowledgements

The authors are grateful to Ruihuang Liu for his helpful discussions regarding the algorithm design.

Conflict of interest

The authors declare that they have no conflicts of interest.

Biographies

Borui Tang is a graduate student at the University of Science and Technology of China. His research mainly focuses on best subset selection and biomedical data analysis.

Junxian Zhu is a Research Fellow at the National University of Singapore. He received his Ph.D. degree in Statistics from Sun Yat-sen University in 2021. His research mainly focuses on best subset selection, GWAS, and noncompliance in randomised clinical trials.

References

[1] Li K-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **1991**, *86* (414): 316–327.
[2] Cook R D. Regression Graphics: Ideas for Studying Regressions through Graphics. New York: Wiley, **1998**.
[3] Li B. Sufficient Dimension Reduction: Methods and Applications with R. New York: Chapman and Hall/CRC, **2018**.
[4] Cook R D. Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **1996**, *91* (435): 983–992.
[5] Carroll R J, Li K C. Measurement error regression with unknown link: Dimension reduction and data visualization. *Journal of the American Statistical Association*, **1992**, *87* (420): 1040–1050.
[6] Chen C-H, Li K-C. Can SIR be as popular as multiple linear regression? *Statistica Sinica*, **1998**, *8* (2): 289–316.
[7] Huang M-Y, Hung H. A review on sliced inverse regression, sufficient dimension reduction, and applications. *Statistica Sinica*, **2022**, *32*: 2297–2314.
[8] Li L, Wen X M, Yu Z. A selective overview of sparse sufficient dimension reduction. *Statistical Theory and Related Fields*, **2020**, *4* (2): 121–133.
[9] Li L, Cook R D, Nachtsheim C J. Model-free variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **2005**, *67* (2): 285–299.
[10] Yin X, Hilafu H. Sequential sufficient dimension reduction for large *p*, small *n* problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **2014**, *77* (4): 879–892.
[11] Yu Z, Dong Y, Shao J. On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *The Annals of*

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

Statistics, **2016**, *44* (6): 2594–2623.
[12] Chen X, Zou C, Cook R D. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, **2010**, *38* (6): 3696–3723.
[13] Ni L, Cook R D, Tsai C-H. A note on shrinkage sliced inverse regression. *Biometrika*, **2005**, *92* (1): 242–247.
[14] Wu Y, Li L. Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica*, **2011**, *2011* (21): 707–730.
[15] Lin Q, Zhao Z, Liu J S. Sparse sliced inverse regression via Lasso. *Journal of the American Statistical Association*, **2019**, *114* (528): 1726–1739.
[16] Tan K M, Wang Z, Liu H, et al. Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **2018**, *80* (5): 1057–1086.
[17] Tan K M, Wang Z, Zhang T, et al. A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, **2018**, *105* (4): 769–782.
[18] Lin Q, Li X, Huang D, et al. On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics*, **2021**, *49* (1): 1–20.
[19] Tan K, Shi L, Yu Z. Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics*, **2020**, *48* (1): 64–85.
[20] Zeng J, Mai Q, Zhang X. Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *Journal of the American Statistical Association*, **2022**, *119*: 343–355.
[21] Zhou K, Zha H, Song L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics. Scottsdale, AZ, USA: PMLR, **2013**: 641–649.
[22] Richard E, Gaiffas S, Vayatis N. Link prediction in graphs with autoregressive features. *The Journal of Machine Learning Research*, **2014**, *15* (1): 565–593.
[23] Guo Y, Zhu Z, Fan J. Best subset selection is robust against design dependence. arXiv: 2007.01478, **2020**.
[24] Zhu J, Canhong Wen C, Zhang H, et al. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, **2020**, *117* (52): 33117–33123.
[25] Wen C, Dong R, Wang X, et al. Best subset selection in reduced rank regression. arXiv: 2211.15889, **2022**.
[26] Zhu J, Zhu J, Tang B, et al. Best-subset selection in generalized linear models: A fast and consistent algorithm via splicing technique. arXiv: 2308.00251, **2023**.
[27] Tang B, Zhu J, Zhu J, et al. A consistent and scalable algorithm for best subset selection in single index models. arXiv: 2309.06230, **2023**.
[28] Cook R D. Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, **2004**, *32* (3): 1062–1092.
[29] Van de Geer S A, Bühlmann P. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, **2009**, *3*: 1360–1392.
[30] Cook R D, Yin X. Theory & methods: special invited paper: dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, **2001**, *43* (2): 147–199.
[31] Ma Y, Zhang X. A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika*, **2015**, *102* (2): 409–420.
[32] Wen C, Zhang A, Quan S, et al. BeSS: an R package for best subset selection in linear, logistic and Cox proportional hazards models. *Journal of Statistical Software*, **2020**, *94*: 1–24.
[33] Zhang Y, Zhu J, Zhu J, et al. A splicing approach to best subset of groups selection. *INFORMS Journal on Computing*, **2023**, *35* (1): 104–119.
[34] Wen C, Li Z, Dong R, et al. Simultaneous dimension reduction and variable selection for multinomial logistic regression. *INFORMS Journal on Computing*, **2023**, *35* (5): 1044–1060.
[35] Zhu J, Wang X, Hu L, et al. abess: A fast best-subset selection library in Python and R. *Journal of Machine Learning Research*, **2022**, *23* (202): 1–7.
[36] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **2007**, *35* (6): 2769–2794.
[37] Sheng W, Yin X. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, **2016**, *25* (1): 91–104.
[38] Hotelling H. Relations between two sets of variates. *Biometrika*, **1936**, *28*: 321–377.
[39] Chung D, Keles S. Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, **2010**, *9* (1): Article17.
[40] Ho Y A H, Xu S, Guo X. Integrative analysis of site-specific parameters with nuisance parameters on the common support. *Statistics in Biosciences*, **2024**: DOI: 10.1007/s12561-024-09428-7.
[41] Jain P, Tewari A, Kar P. On iterative hard thresholding methods for high-dimensional M-estimation. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Volume 1. Cambridge, MA, USA: MIT Press, **2014**: 685–693.
[42] Ghojogh B, Karray F, Crowley M. Eigenvalue and generalized eigenvalue problems: Tutorial. arXiv: 1903.11240, **2023**.

Appendix

A.1 Technical proofs

Proof of Proposition 1. Note that for any matrix $\boldsymbol{X} \in \mathbb{R}_{m\times n}$, we have

$$\boldsymbol{X}\boldsymbol{X}^{\top} = \begin{pmatrix} \boldsymbol{X}_{1,\cdot} \\ \boldsymbol{X}_{2,\cdot} \\ \vdots \\ \boldsymbol{X}_{m,\cdot} \end{pmatrix} \left(\boldsymbol{X}_{1,\cdot}, \boldsymbol{X}_{2,\cdot}, \cdots, \boldsymbol{X}_{m,\cdot} \right) = \{ \boldsymbol{X}_{i,\cdot} \boldsymbol{X}_{j,\cdot}^{\top} \}_{i,j}.$$

It follows that

$$\boldsymbol{B}^{\star}\boldsymbol{B}^{\star\top} - \boldsymbol{B}^{\dagger}\boldsymbol{B}^{\dagger\top} = \{ \boldsymbol{B}_{i,\cdot}^{\star}\boldsymbol{B}_{j,\cdot}^{\star\top} - \boldsymbol{B}_{i,\cdot}^{\dagger}\boldsymbol{B}_{j,\cdot}^{\dagger\top} \}_{i,j}.$$

Suppose that $\mathrm{supp}(\boldsymbol{B}^{\dagger}) \not\subseteq \mathrm{supp}(\boldsymbol{B}^{\star})$, there exist $j \in \{1, \cdots, p\}$ such that $j \in \mathrm{supp}(\boldsymbol{B}^{\dagger})$ and $j \notin \mathrm{supp}(\boldsymbol{B}^{\star})$. Thus,

$$\rho(\boldsymbol{B}^{\dagger}, \boldsymbol{B}^{\star}) \geq (\boldsymbol{B}_{j,\cdot}^{\star}\boldsymbol{B}_{j,\cdot}^{\star\top} - \boldsymbol{B}_{j,\cdot}^{\dagger}\boldsymbol{B}_{j,\cdot}^{\dagger\top})^2 = \|\boldsymbol{B}_{j,\cdot}^{\dagger}\|_2^4.$$

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

This contradicts our condition.

Proof of Lemma 1. Define function

$$\boldsymbol{C}^{\circ}(i,\boldsymbol{v}) = \left(\boldsymbol{C}_{1,\cdot}^{\circ\top},\cdots,\boldsymbol{C}_{i-1,\cdot}^{\circ\top},\boldsymbol{v}^{\top},\boldsymbol{C}_{i+1,\cdot}^{\circ\top},\cdots,\boldsymbol{C}_{p,\cdot}^{\circ\top}\right)^{\top}.$$

By definition,

$$\begin{aligned}\boldsymbol{C}_{i\cdot}^{\circ} =& \operatorname*{argmin}_{\boldsymbol{v}} \tfrac{\rho}{2}\|\boldsymbol{C}^{\circ}(i,j,\boldsymbol{v}) - \boldsymbol{B}^{\circ}\|_F^2 + \langle \boldsymbol{D}^{\circ}, \boldsymbol{C}^{\circ}(i,j,\boldsymbol{v}) - \boldsymbol{B}^{\circ}\rangle + \mu^{\circ}\|\boldsymbol{C}^{\circ}(i,j,\boldsymbol{v})\|_{0,2} = \\ & \operatorname*{argmin}_{\boldsymbol{v}} \tfrac{\rho}{2}\|\boldsymbol{v} - \boldsymbol{B}_{i,\cdot}^{\circ}\|_2^2 + \langle D_{i,\cdot}^{\circ}, \boldsymbol{v} - \boldsymbol{B}_{i,\cdot}^{\circ}\rangle + \mu^{\circ}\|\boldsymbol{v}\|_0.\end{aligned}$$

Differentiating with respect to $\boldsymbol{v}$, by simple calculation and the notion of sub-gradient, we can easily obtain the desired equation.

Proof of Proposition 3. For $T < s = |\mathcal{A}^m|$, we have

$$\min_{j\in S_{T,2}^m} \frac{1}{\kappa^2}\|\boldsymbol{D}_{i,\cdot}^m\|_2^2 = \min_{j\in S_{T,2}^m} \frac{1}{\kappa^2}\|\boldsymbol{\beta}_{i,\cdot}^m + \frac{1}{\rho}\boldsymbol{D}_{i,\cdot}^m\|_2^2 \geq \max_{i\in\mathcal{A}^m}\|\boldsymbol{\beta}_{i,\cdot}^m + \frac{1}{\rho}\boldsymbol{D}_{i,\cdot}^m\|_2^2 = \max_{i\in\mathcal{A}^m}\|\boldsymbol{\beta}_{i,\cdot}^m\|_2^2 \geq \max_{i\in S_{T,1}^m}\|\boldsymbol{\beta}_{i,\cdot}^m\|_2^2.$$

Thus, the corresponding range of ρ is

$$\rho \leq \frac{\min_{i\in S_{T,2}^m}\|\boldsymbol{D}_{i,\cdot}^m\|_2}{\max_{i\in S_{T,1}^m}\|\boldsymbol{B}_{i,\cdot}^m\|_2}. \tag{A1}$$

Similarly, given $T+1$, the range of κ is

$$\rho \leq \frac{\min_{i\in S_{T+1,2}^m}\|\boldsymbol{D}_{i,\cdot}^m\|_2}{\max_{i\in S_{T+1,1}^m}\|\boldsymbol{B}_{i,\cdot}^m\|_2}. \tag{A2}$$

Note that $S_{T,1}^m \subseteq S_{T+1,1}^m$ and $S_{T,2}^m \subseteq S_{T+1,2}^m$. Given T, ρ lies in the difference set between the above two ranges. Therefore, given $T < |\mathcal{A}^m|$, the corresponding range of ρ is

$$\rho \in \left(\frac{\min_{i\in S_{T+1,2}^{m+1}}\|\boldsymbol{D}_{i,\cdot}^m\|}{\max_{i\in S_{T,1}^{m+1}}\|\boldsymbol{B}_{i,\cdot}^m\|}, \frac{\min_{i\in S_{T+1,2}^{m+1}}\|\boldsymbol{D}_{i,\cdot}^m\|}{\max_{i\in S_{T,1}^{m+1}}\|\boldsymbol{B}_{i,\cdot}^m\|}\right].$$

Given $T = s = |\mathcal{A}^m|$, $S_{T,1}^m = \mathcal{A}^m$. The corresponding range of ρ is

$$\rho \in \left(0, \frac{\min_{i\in S_{T,2}^{m+1}}\|\boldsymbol{D}_{i,\cdot}^m\|}{\max_{i\in S_{T,1}^{m+1}}\|\boldsymbol{B}_{i,\cdot}^m\|}\right].$$

A.2 Additional simulation results

A.2.1 Simulations under single index models

This section presents simulation results for single index models with one-dimensional central subspaces. Specifically, we consider the following two models:

(c) $Y = X^{\top}\boldsymbol{\beta} + \epsilon$.

(d) $Y = \exp(X\boldsymbol{\beta}) + \epsilon$.

Model (c) is a linear model, and model (d) is a single index model with an exponential transformation. We set the first eight elements of β to 0.5 and the others to 0. All other settings are the same as in Section 3. The results are recorded in Figs. A1 and A2.

As shown in Figs. A1 and A2, splicing-SIR and Relaxed Natural SSIR can recover the true support with a high probability, while LASSO-SIR and SEAS-SIR always fail. Moreover, Splicing-SIR leads to more accurate estimators — the canonical correlations are higher.

A.2.2 Simulations in high-dimensional scenarios

The splicing-SIR algorithm is capable of handling high-dimensional data. Before presenting the simulation results, we illustrate a computational technique in such scenarios. The initialization procedure, detailed in Section 2.4, involves generalized eigenvalue decomposition for $(\widehat{\boldsymbol{M}},\widehat{\boldsymbol{\Sigma}})$. However, high-dimensional data typically lead to a singular $\widehat{\boldsymbol{\Sigma}}$.To address this issue and enhance computational stability, we follow common practice by modifying $\widehat{\boldsymbol{\Sigma}}$ to $\widehat{\boldsymbol{\Sigma}} + u\boldsymbol{I}_p$, where u is a small constant[42]. We set u to 10^{-8} in our implementation.

We examine the performance of the splicing-SIR, SEAS-SIR, and LASSO-SIR algorithms under varying magnitudes of noise. The Relaxed Natural SSIR is excluded from this investigation due to its computational infeasibility with high-dimensional data. Specifically, we generate the noise variable from the normal distribution $\mathcal{N}(0,\sigma^2)$, where σ is set as $0, 0.1, 1$, and 10, respectively. The design matrix setting is the same as that described in Section 3. For the coefficients $\boldsymbol{\beta}$, we assigned a value of 0.5 to four elements of the first column and another four elements of the second column, and we set all other elements to 0. Finally, the

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

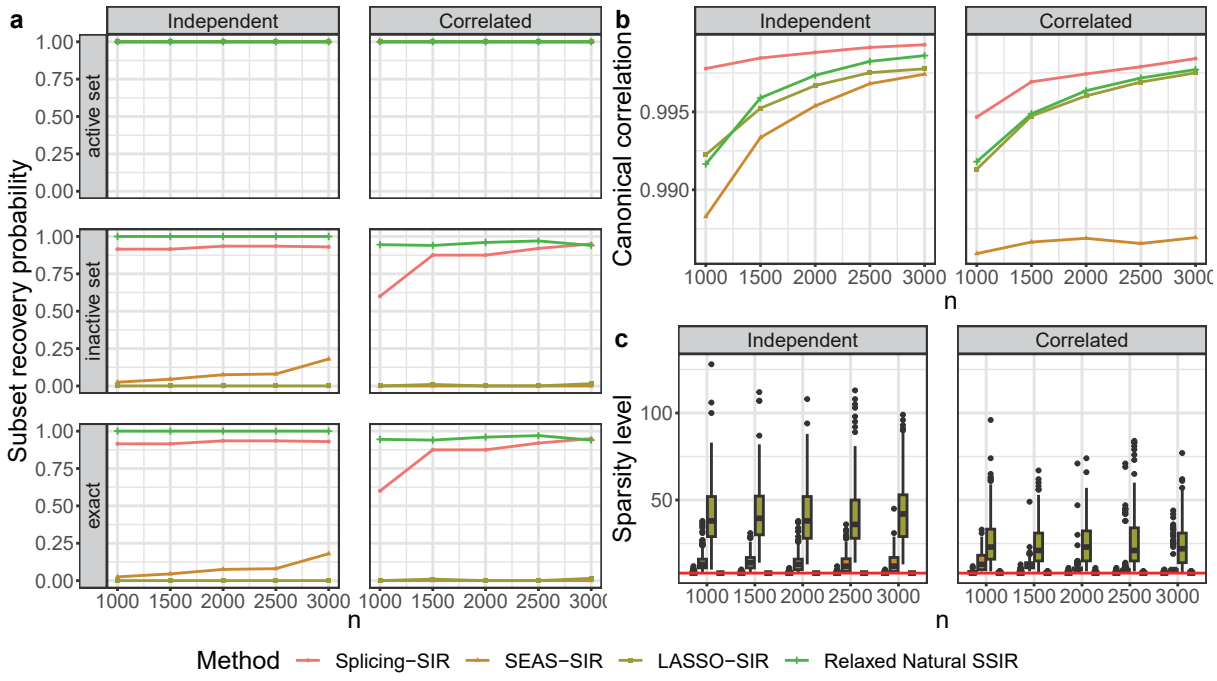


Fig. A1. Statistical performance under Model (c). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

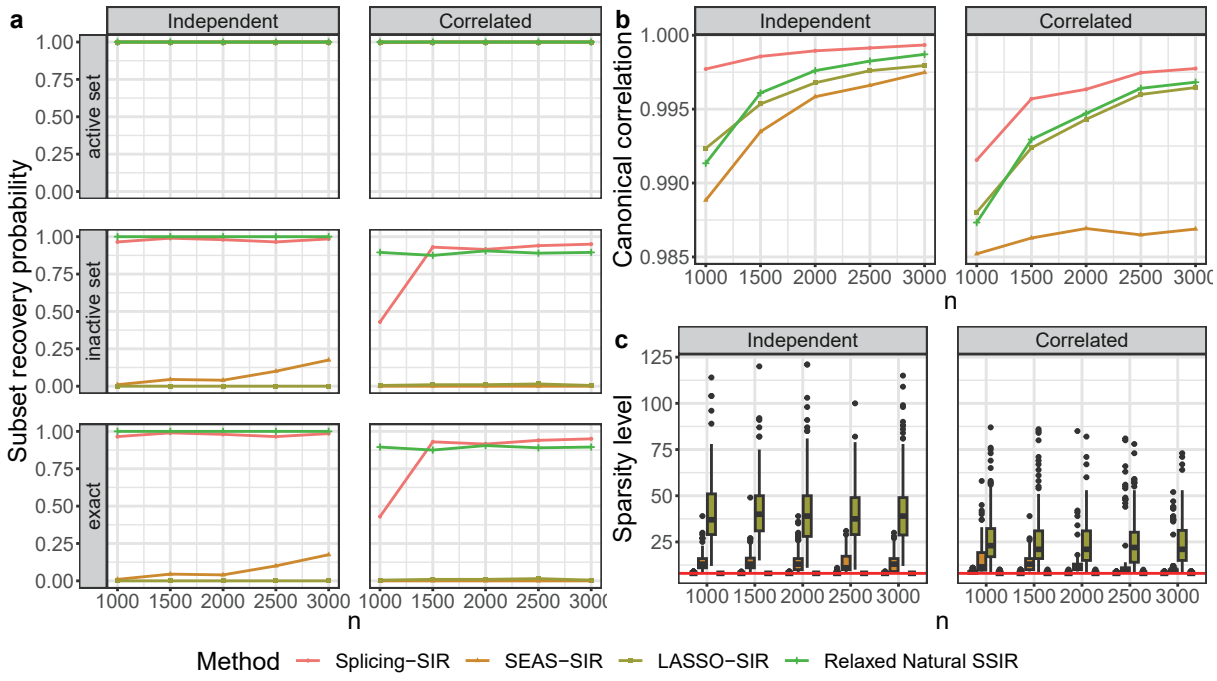


Fig. A2. Statistical performance under Model (d). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

response values are generated according to Models (a) and (b), respectively. All experiments are conducted on 100 synthetic datasets, each with a fixed sample size of $n = 1000$, and a dimension of $p = 4000$. The simulation results are presented in Figs. A3 and A4.

As shown in the figures, the exact subset recovery probability of the splicing-SIR algorithm increases as the magnitude of the noise decreases. In the independent design scenario, splicing-SIR recovers the true subset with probabilities exceeding 0.85 when no noise exists. In contrast, LASSO-SIR and SEAS-SIR fail in most cases. In the correlated design scenario, although all methods tend to mistakenly include irrelevant variables, splicing-SIR performs relatively better. Moreover, it yields more accurate estimators, as evidenced by higher canonical correlations.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503

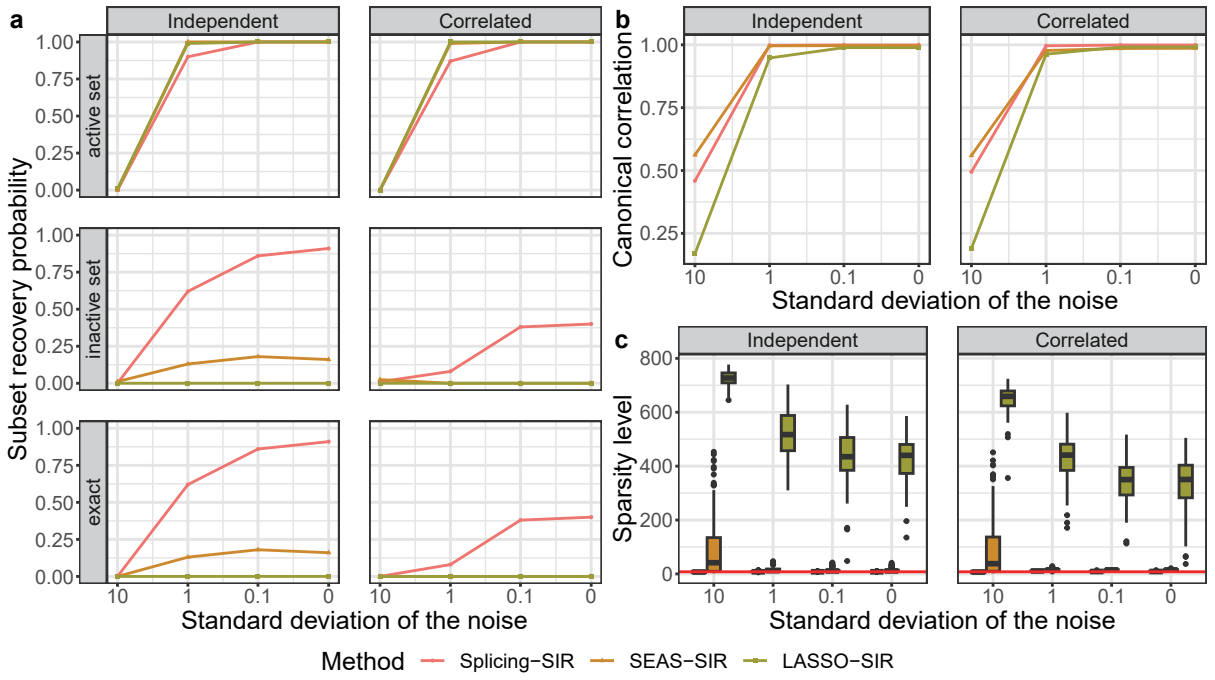


Fig. A3. Statistical performance in high-dimensional scenarios under Model (a). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

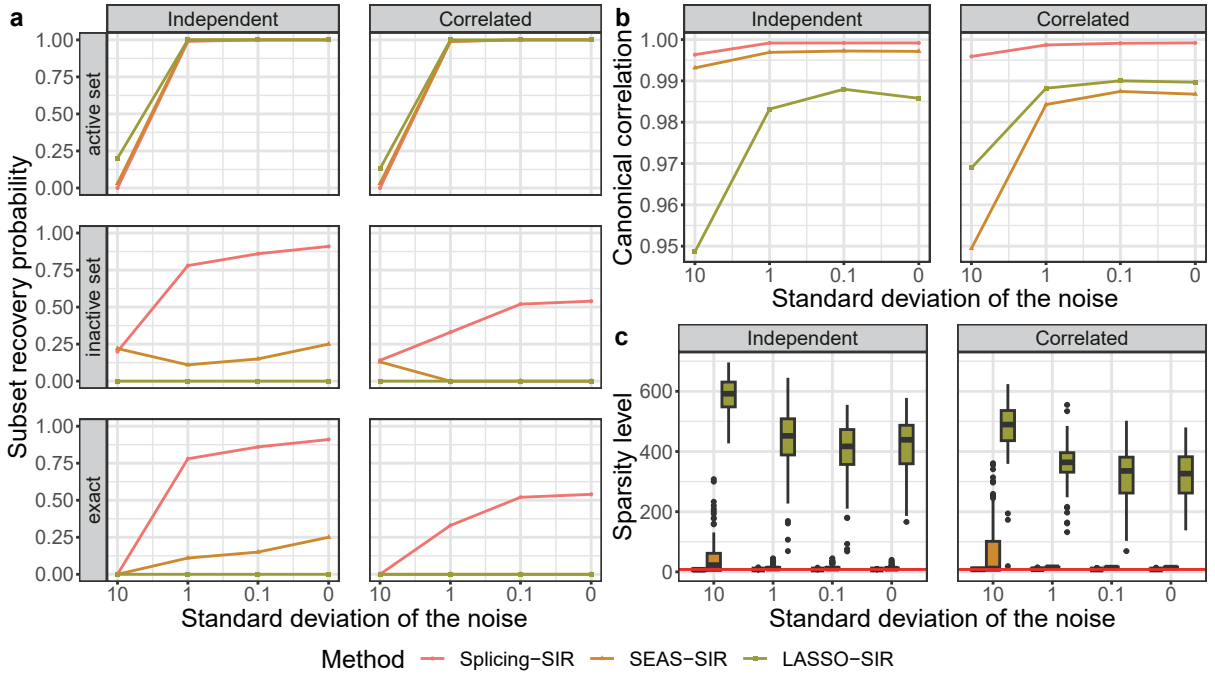


Fig. A4. Statistical performance in high-dimensional scenarios under Model (b). (a) Subset recovery probability for the active set, inactive set, and both sets. (b) Median of canonical correlations. (c) Sparsity level of the selected model.

CSTR: 32290.14.JUSTC-2024-0034
DOI: 10.52396/JUSTC-2024-0034
JUSTC, **2025**, 55(5): 0503