Journal of Multivariate Analysis 112 (2012) 248–255

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Note(s)

On the consistency of coordinate-independent sparse estimation with BIC☆

Changliang Zou [a,b,*], Xin Chen [c]

[a] LPMC, Nankai University, Tianjin, China

[b] Department of Statistics, Nankai University, Tianjin, China

[c] Department of Statistics and Applied Probability, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 2 November 2011
Available online 7 May 2012

AMS 2000 subject classifications:
62H20
62J07

Keywords:
BIC
Central subspace
Consistency
Coordinate-independent
Sufficient dimension reduction
Variable selection

ABSTRACT

Chen et al. (2010) [1] propose a unified method – coordinate-independent sparse estimation (CISE) – that is able to simultaneously achieve sparse sufficient dimension reduction and screen out irrelevant and redundant variables efficiently. However, its attractive features depend on the appropriate choice of the tuning parameter. In this note, we re-examine the Bayesian information criterion (BIC) in sufficient dimension reduction and provide a heuristic derivation. Furthermore, the CISE with BIC is shown to be able to identify the true model consistently.

1. Introduction

Consider the regression of a univariate response y on p random predictors $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, with the general goal of inferring about the conditional distribution of $y|\mathbf{x}$. Sufficient dimension reduction (SDR) in regression, which reduces the dimension by replacing original predictors with a minimal set of their linear combinations without loss of information, is very helpful when the number of predictors is large [2]. Many SDR methods, including both moment-based and model-based, can be formulated as a generalized eigenvalue problem in the following form [7,1]

$$\mathbf{M}_n\boldsymbol{\delta}_{ni} = \lambda_{ni}\mathbf{N}_n\boldsymbol{\delta}_{ni}, \quad \text{for } i = 1, \ldots, p, \tag{1}$$

where $\mathbf{M}_n \geq 0$ is a method-specific symmetric kernel matrix; For example, the population versions of $\mathbf{M}_n$ for principle component analysis and sliced inverse regression (SIR) are $\mathrm{cov}(\mathbf{x})$ and $\mathrm{cov}[E\{\mathbf{x} - E(\mathbf{x})|y\}]$, respectively (see Table 1 in [1] for more examples); $\mathbf{N}_n > 0$ is symmetric, often taking the form of the sample covariance matrix $\boldsymbol{\Sigma}_n$ of $\mathbf{x}$; $\boldsymbol{\delta}_{n1}, \ldots, \boldsymbol{\delta}_{np}$ are eigenvectors such that $\boldsymbol{\delta}_{ni}^T\mathbf{N}_n\boldsymbol{\delta}_{nj} = 1$ if $i = j$ and 0 if $i \neq j$, and $\lambda_{n1} \geq \cdots \geq \lambda_{np}$ are the corresponding eigenvalues. We use the subscript "n" to indicate that $\boldsymbol{\Sigma}_n$, $\mathbf{M}_n$, $\mathbf{N}_n$ and λ_{ni} are the sample versions of the corresponding population analogs $\boldsymbol{\Sigma}$, $\mathbf{M}$, $\mathbf{N}$ and λ_i. Under certain conditions that are usually imposed only on the marginal distribution of $\mathbf{x}$, the eigenvectors

☆ This research was supported by the NNSF of China Grants 11001138, 11071128, 11131002, 11101306, the RFDP of China Grant 20110031110002, National University of Singapore Research Grants and National Center for Theoretical Sciences, Math Division.

* Correspondence to: School of Mathematical Sciences, Nankai University, Tianjin, 300071, China.
E-mail address: chlzou@yahoo.com.cn (C. Zou).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2012.04.014

$\{\boldsymbol{\delta}_{n1}, \ldots, \boldsymbol{\delta}_{nd}\}$ that correspond to the nonzero eigenvalues $\lambda_{n1} \geq \cdots \geq \lambda_{nd}$ form a consistent estimator of a basis for the central subspace. The dimension d, usually far less than p, is assumed to be known in this article. We also assume throughout that $n > p$.

Based on (1), the standard sufficient dimension reduction estimation (SSDRE) can be expressed as [7]

$$\widehat{\mathbf{V}} = \underset{\mathbf{V}\in\mathbb{R}^{p\times d}}{\arg\min} -\mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}), \quad \text{subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d. \tag{2}$$

The SSDRE often suffers because the estimated linear combinations usually consist of all original predictors, making it difficult to interpret [10,7]. By using the coordinate independent penalty function, Chen et al. [1] proposed a unified method – coordinate-independent sparse estimation (CISE) – that can simultaneously achieve sparse sufficient dimension reduction and screen out irrelevant and redundant variables efficiently. Formally, the CISE is defined by

$$\tilde{\mathbf{V}} = \underset{\mathbf{V}\in\mathbb{R}^{p\times d}}{\arg\min}\{-\mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}) + \rho(\mathbf{V})\}, \quad \text{subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d, \tag{3}$$

where $\rho(\mathbf{V})$ is defined by $\rho(\mathbf{V}) = \sum_{i=1}^{p}\theta_i\|\mathbf{v}_i\|_2$, $\mathbf{v}_i$ is the ith row of $\mathbf{V}$ and $\theta_i \geq 0$ are penalty parameters. This CISE is subspace oriented and thus finding $\tilde{\mathbf{V}}$ results in a Grassmann manifold optimization problem. A fast algorithm is suggested by Chen et al. [1]. Under mild conditions, they also established the oracle property of CISE in the sense that it would perform asymptotically as well as if the true irrelevant predictors were known.

The above features of the CISE method rely on the proper choice of tuning parameters, or called regularization parameter, which is usually selected by some criteria, such as cross-validation, C_p and generalized cross-validation [5,19]. Chen et al. [1] recommended using a Bayesian information criterion (BIC; [14]) to determine the tuning parameters by mimicking the classical BIC in the context of multiple regression. In this paper, we revisit the definition of BIC for the SDR methods and give a heuristic but formal Bayesian derivation. Then we prove that the CISE with the BIC identifies the true model consistently.

2. The BIC in SDR

We need the following notation and definition for ease of exposition. Define the Stiefel manifold $\mathrm{St}(p, d)$ as $\mathrm{St}(p, d) = \{\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d} : \boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d\}$. Denotes $\lfloor\boldsymbol{\Gamma}\rfloor$ as the subspace spanned by the columns of $\boldsymbol{\Gamma}$, then $\lfloor\boldsymbol{\Gamma}\rfloor \in \mathrm{Gr}(p, d)$ where $\mathrm{Gr}(p, d)$ stands for the Grassmann manifold. Define the matrix norm $\|\mathbf{V}\|_t = \sqrt{\mathrm{tr}(\mathbf{V}^T\mathbf{V})}$. The projection operator $R : \mathbb{R}^{p\times d} \to \mathrm{St}(p, d)$ onto the Stiefel manifold $\mathrm{St}(p, d)$ is defined to be $R(\boldsymbol{\Gamma}) = \arg\min_{\mathbf{W}\in\mathrm{St}(p,d)}\|\boldsymbol{\Gamma} - \mathbf{W}\|_t^2$. The tangent space $T_{\boldsymbol{\Gamma}}(p, d)$ of $\boldsymbol{\Gamma} \in \mathrm{St}(p, d)$ is defined by

$$T_{\boldsymbol{\Gamma}}(p, d) = \left\{\mathbf{Z} \in \mathbb{R}^{p\times d} : \mathbf{Z} = \boldsymbol{\Gamma}\mathbf{A} + \boldsymbol{\Gamma}_\perp\mathbf{B}, \mathbf{A} \in \mathbb{R}^{d\times d}, \mathbf{A} + \mathbf{A}^T = \mathbf{0}, \mathbf{B} \in \mathbb{R}^{(p-d)\times d}\right\}, \tag{4}$$

where $\boldsymbol{\Gamma}_\perp \in \mathbb{R}^{p\times(p-d)}$ is the complement of $\boldsymbol{\Gamma}$ which satisfies $[\boldsymbol{\Gamma}\ \boldsymbol{\Gamma}_\perp]^T[\boldsymbol{\Gamma}\ \boldsymbol{\Gamma}_\perp] = \mathbf{I}_p$.

2.1. Definition

Let s be a subset of $\{1, \ldots, p\}$. Denote by $\mathbf{V}_s$ the parameter $\mathbf{V}$ with those rows outside s being $\mathbf{0}$, that is, $\forall i \notin s$, $\mathbf{v}_i = \mathbf{0}$. Denote by

$$\widehat{\mathbf{V}}_s = \underset{\{\mathbf{V}_s\in\mathbb{R}^{p\times d}:\mathbf{V}_s^T\mathbf{N}_n\mathbf{V}_s=\mathbf{I}_d\}}{\arg\min} -\mathrm{tr}(\mathbf{V}_s^T\mathbf{M}_n\mathbf{V}_s)$$

the SSDRE given model s, and by df_s the effective number of parameters. The BIC criterion in SDR methods is presented in the following proposition.

Proposition 1. *The generalized BIC in SDR methods chooses the model that yields the smallest value of*

$$-\mathrm{tr}(\widehat{\mathbf{V}}_s^T\mathbf{M}_n\widehat{\mathbf{V}}_s) + \mathrm{df}_s \cdot \log(n)/n, \tag{5}$$

where $\mathrm{df}_s = (p_s - d)\cdot d$ *and* p_s *is the number of non-zero rows of* $\widehat{\mathbf{V}}_s$.

In fact, the degree of freedom $(p_s - d)\cdot d$ comes from the fact that we need $(p_s - d)\cdot d$ parameters to describe a d-dimensional Grassmann manifold in $\mathbb{R}^{p_s}$ [4]. In what follows, without loss of generality, we assume that only the first $q < p$ predictors are relevant to the regression. Next we will give a formal derivation of this criterion.

2.2. A heuristic Bayesian derivation

In the Bayesian framework, model comparison is based on posterior probabilities. Consider a candidate model $s \in \mathcal{S}$ where $\mathcal{S}$ is the model space under consideration. Assume that model s has a prior probability $\pi(s)$, and the prior density of its parameter α_s is $\pi(\alpha_s|s)$. Then the posterior probability of model s given data D satisfies

$$p(s|D) \propto \pi(s)\int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s.$$

Under the Bayes paradigm, a model s^* that maximizes the posterior probability is selected, say $s^* = \arg\max_{s\in\mathcal{S}} \pi(s) \int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s$. In practice, an implicit assumption underlying is that the candidate models are equally likely so that $\pi(s)$ is constant over $\mathcal{S}$. Consequently, model assessment mainly depends on the integral term $\int p(D|\alpha_s, s)\pi(\alpha_s|s)d\alpha_s$ which is usually referred to as marginal likelihood for model s.

The classical Schwarz's BIC is an approximation to the logarithm of the marginal likelihood, and there is a similar heuristic derivation for the generalized BIC in SDR. We consider a pseudo-likelihood [16]

$$L(D|\mathbf{V}_s, s) \propto \exp\{-ng(\mathbf{V}_s; \mathbf{M}_n)/2\}, \tag{6}$$

where $g(\mathbf{V}_s; \mathbf{M}_n)$ is the generalized eigenvalue loss function in (5) corresponding to model s. The main motivation for using (6) as a pseudo-likelihood is two-fold: on one hand, minimizing $g(\mathbf{V}_s; \mathbf{M}_n)$ gives the SSDRE which works like maximizing a log-likelihood function (cf., [7]); on the other hand, $g(\mathbf{V}_s; \mathbf{M}_n)/2$ happens to be the log-likelihood by ignoring some constants with respect to $\mathbf{V}_s$, under the PFC-model and the normality assumption [3]. Note that $g(\mathbf{V}_s; \mathbf{M}_n) = -\mathrm{tr}(\mathbf{V}_{s(s)}^T \mathbf{M}_{n(s)} \mathbf{V}_{s(s)})$ in which $\mathbf{V}_{s(s)}^T \mathbf{N}_{(s)} \mathbf{V}_{s(s)} = \mathbf{I}_d$. Here, given a $p \times d$ matrix $\mathbf{K}$, $\mathbf{K}_{(s)}$ indicates the sub-matrix which consists of all rows of $\mathbf{K}$ whose indices are in s. If $\mathbf{K}$ is $p \times p$ then the notation indicates the sub-matrix which consists of all rows and columns of $\mathbf{K}$ whose indices are in s.

Next, we consider approximate the integral pseudo-likelihood,

$$\int \exp\{-ng(\mathbf{V}_{s(s)}; \mathbf{M}_{n(s)})/2\}\pi(\mathbf{V}_{s(s)}|s)d\mathbf{V}_{s(s)}$$

by using the Laplace method [15]. As we know, the basic idea of Laplace approximation is that in large samples, the integral is largely determined by the value of the integrand in a region close to $\widehat{\mathbf{V}}_\pi \in \mathbb{R}^{p_s\times d}$, the value of $\mathbf{V}_{s(s)}$ that maximizes $\tilde{g}(\mathbf{V}_{s(s)}; \mathbf{M}_{n(s)}) = -ng(\mathbf{V}_{s(s)}; \mathbf{M}_{n(s)}) + \log(\pi(\mathbf{V}_{s(s)}|s))$ subject to $\mathbf{V}_{s(s)}^T \mathbf{N}_{n(s)} \mathbf{V}_{s(s)} = \mathbf{I}_d$. For Schwarz's BIC, the Laplace approximation is performed by a second-order Taylor expansion of $\tilde{g}(\mathbf{V}_{s(s)}; \mathbf{M}_{n(s)})$ around $\widehat{\mathbf{V}}_\pi$, but the same approach is not directly feasible here because $\mathbf{V}_{s(s)}$ must satisfy the constraint $\mathbf{V}_{s(s)}^T \mathbf{N}_{n(s)} \mathbf{V}_{s(s)} = \mathbf{I}_d$. As mentioned before, the priors used for model s are typically the unit information prior [12]. Hence, in such cases, we have $\widehat{\mathbf{V}}_\pi \approx \widehat{\mathbf{V}}(s)$, the SSDRE defined in (1) with $\mathbf{M}_{n(s)}$ and $\mathbf{N}_{n(s)}$ instead. We will give an approximation of $g(\mathbf{V}_{s(s)}; \mathbf{M}_{n(s)})$ in a small neighborhood of $\widehat{\mathbf{V}}(s)$.

Following Proposition 2 in [1], we work under the following equivalent unitary constraints optimization problems which will facilitate our exposition, i.e., $\widehat{\mathbf{V}}(s) = \mathbf{N}_{n(s)}^{-1/2}\widehat{\mathbf{\Gamma}}_s$, where

$$\widehat{\mathbf{\Gamma}}_s = \underset{\mathbf{\Gamma}\in\mathbb{R}^{p_s\times d}}{\arg\min} -\mathrm{tr}(\mathbf{\Gamma}^T \mathbf{G}_n(s)\mathbf{\Gamma}), \quad \text{subject to } \mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d,$$

and $\mathbf{G}_n(s) = \mathbf{N}_{n(s)}^{-1/2}\mathbf{M}_{n(s)}\mathbf{N}_{n(s)}^{-1/2}$. For an arbitrary matrix $\mathbf{W} \in \mathbb{R}^{p_s\times d}$ and scalar $\delta \in \mathbb{R}$, the perturbed point around $\lfloor\widehat{\mathbf{\Gamma}}_s\rfloor$ in the Grassmann manifold can be expressed by $\lfloor R(\widehat{\mathbf{\Gamma}}_s + \mathbf{W})\rfloor$, where $\mathbf{W}$ can be uniquely decomposed as [8, Lemma 8]

$$\mathbf{W} = \widehat{\mathbf{\Gamma}}_s\mathbf{A} + \widehat{\mathbf{\Gamma}}_{s\perp}\mathbf{B} + \widehat{\mathbf{\Gamma}}_s\mathbf{C},$$

in which $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a skew-symmetric matrix, $\mathbf{B} \in \mathbb{R}^{(p_s-d)\times d}$ is an arbitrary matrix, and $\mathbf{C} \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Let $\mathbf{Z} = \widehat{\mathbf{\Gamma}}_s\mathbf{A} + \widehat{\mathbf{\Gamma}}_{s\perp}\mathbf{B}$. Obviously, $\mathbf{Z} \in T_{\widehat{\mathbf{\Gamma}}_s}(p_s, d)$. By Chen et al. [1], the movement from $\lfloor\widehat{\mathbf{\Gamma}}_s\rfloor$ in the near neighborhood only depends on $\widehat{\mathbf{\Gamma}}_{s\perp}\mathbf{B}$. In other words, to obtain an expansion around $\widehat{\mathbf{\Gamma}}_s$, it suffices to only consider perturbed points like $R(\widehat{\mathbf{\Gamma}}_s + \delta\mathbf{Z})$, where $\|\mathbf{B}\|_t = C$ for some given C. Let $\lambda_{n1}(s) \geq \cdots \geq \lambda_{np_s}(s) \geq 0$ be the eigenvalues of $\mathbf{G}_n(s)$, $\mathbf{\Lambda}_n(s) = \mathrm{diag}\{\mathbf{\Lambda}_{n1}(s), \mathbf{\Lambda}_{n2}(s)\}$ with $\mathbf{\Lambda}_{n1}(s) = \mathrm{diag}\{\lambda_{n1}(s), \ldots, \lambda_{nd}(s)\}$ and $\mathbf{\Lambda}_{n2}(s) = \mathrm{diag}\{\lambda_{nd+1}(s), \ldots, \lambda_{np_s}(s)\}$. By using Lemma 1-(ii) in [1], we have

$$\begin{aligned}
&g(R(\widehat{\mathbf{\Gamma}}_s + \mathbf{Z}); \mathbf{G}_n(s)) \qquad (7)\\
&\quad\approx g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) - 2\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n(s)\widehat{\mathbf{\Gamma}}_s) - \mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n(s)\mathbf{Z}) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\widehat{\mathbf{\Gamma}}_s^T\mathbf{G}_n(s)\widehat{\mathbf{\Gamma}}_s)\\
&\quad= g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) - 2\mathrm{tr}(\mathbf{Z}^T\widehat{\mathbf{\Gamma}}_s\mathbf{\Lambda}_{n1}(s)) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Lambda}_{n1}(s)) - \mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n(s)\mathbf{Z})\\
&\quad= g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) + \mathrm{tr}(\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}_{n1}(s)) + \mathrm{tr}(\mathbf{B}^T\mathbf{B}\mathbf{\Lambda}_{n1}(s)) - \mathrm{tr}(\mathbf{B}\mathbf{B}^T\mathbf{\Lambda}_{n2}(s)) - \mathrm{tr}(\mathbf{A}\mathbf{A}^T\mathbf{\Lambda}_{n1}(s))\\
&\quad= g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) + \mathrm{tr}(\mathbf{B}\mathbf{\Lambda}_{n1}(s)\mathbf{B}^T) - \mathrm{tr}(\mathbf{B}^T\mathbf{\Lambda}_{n2}(s)\mathbf{B}) := g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) + nh(\mathbf{B}), \qquad (8)
\end{aligned}$$

where the first equality follows from Proposition 2 in [1], the second equality holds because $\mathrm{tr}(\mathbf{Z}^T\widehat{\mathbf{\Gamma}}_s\mathbf{\Lambda}_{n1}(s)) = 0$ by Lemma 1-(i) in [1], and the last equality comes from the fact $\mathrm{tr}(\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}_{n1}(s)) - \mathrm{tr}(\mathbf{A}\mathbf{A}^T\mathbf{\Lambda}_{n1}(s)) = 0$ because $\mathbf{A}$ is skew-symmetric. Consequently, in a small neighborhood around $\widehat{\mathbf{\Gamma}}_s$, say $\widehat{\mathbf{\Gamma}}_s + \mathbf{Z}$, the loss function $g(\mathbf{\Gamma})$ can be approximated by the function $g(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s)) + nh(\mathbf{B})$.

It is easily seen that $h(\mathbf{B})$ can be equivalently represented as the vector form $h(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\tilde{\mathbf{\Lambda}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = \mathrm{Vec}(\mathbf{B})$ and $\tilde{\mathbf{\Lambda}}$ is a $[(p_s - d)d]$-dimensional symmetric matrix only depending on $\mathbf{\Lambda}_{n1}(s)$ and $\mathbf{\Lambda}_{n2}(s)$. Note that if $\lambda_{nd}(s) > \lambda_{nd+1}(s)$, we will have

$$h(\mathbf{B}) \geq (\lambda_{nd}(s) - \lambda_{nd+1}(s))\|\mathbf{B}\|_t^2 > 0, \quad \text{for any } \mathbf{B} \neq \mathbf{0},$$

which follows from basic properties of the trace operator for a semi-positive definite matrix. Thus, it can be concluded that $\tilde{\mathbf{\Lambda}}$ is a positive definite matrix. Now, applying Laplace approximation we obtain

$$\begin{aligned}\int \exp\{-ng(\mathbf{\Gamma}; \mathbf{G}_n(s))/2\}\pi(\widehat{\mathbf{\Gamma}}|s)d\mathbf{\Gamma} &\approx \exp\{-ng(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s))/2 + \log(\pi(\widehat{\mathbf{\Gamma}}|s))\}\int \exp\{-nh(\mathbf{B})/2\}d\mathbf{B}\\ &= \exp\{-ng(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s))/2 + \log(\pi(\widehat{\mathbf{\Gamma}}|s))\}\int \exp\{-n\boldsymbol{\beta}^T\tilde{\mathbf{\Lambda}}\boldsymbol{\beta}/2\}d\boldsymbol{\beta}\\ &= \exp\{-ng(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s))/2 + \log(\pi(\widehat{\mathbf{\Gamma}}|s))\}(2\pi)^{(p_s-d)d/2}|n\tilde{\mathbf{\Lambda}}|^{-1/2}.\end{aligned}$$

We immediately obtain

$$\begin{aligned}&\log\left[\int \exp\{-ng(\mathbf{\Gamma}; \mathbf{G}_n(s))/2\}\pi(\widehat{\mathbf{\Gamma}}|s)d\mathbf{\Gamma}\right]\\ &\quad\approx -ng(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s))/2 + \log(\pi(\widehat{\mathbf{\Gamma}}|s)) + [(p_s-d)d/2]\log(2\pi) - (1/2)\log|n\tilde{\mathbf{\Lambda}}|\\ &\quad= -ng(\widehat{\mathbf{\Gamma}}_s; \mathbf{G}_n(s))/2 - [(p_s-d)d/2]\log n + O(1).\end{aligned}$$

If we ignore terms of $O(1)$ order, finding the model that gives the highest posterior probability based on the pseudo-likelihood (6) leads to minimizing the generalized BIC defined in (5).

2.3. Consistency of the CISE with BIC

Schwarz's BIC is consistent in the sense that it selects the true model with probability approaching one if such a true model is in the class of candidate models. In practice, when p is large, we cannot afford to calculate the BIC values (5) for all possible s. Instead, we prefer to combine this criterion with the coordinate-independent penalized technique developed by Chen et al. [1], which leads to

$$\mathrm{BIC}_{\boldsymbol{\theta}} = -\mathrm{tr}(\tilde{\mathbf{V}}_{\boldsymbol{\theta}}^T\mathbf{M}_n\tilde{\mathbf{V}}_{\boldsymbol{\theta}}) + [(p_{\boldsymbol{\theta}} - d)d]\cdot\log n/n, \tag{9}$$

where $\tilde{\mathbf{V}}_{\boldsymbol{\theta}}$ denotes the solution of (3) for $\mathbf{V}$ given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and $p_{\boldsymbol{\theta}}$ denotes the number of non-zero rows of $\tilde{\mathbf{V}}_{\boldsymbol{\theta}}$. The estimator $\tilde{\mathbf{V}}_{\boldsymbol{\theta}}$ naturally defines a candidate model $s_{\boldsymbol{\theta}} = \{i : \tilde{\mathbf{v}}_{\boldsymbol{\theta}i} \neq \mathbf{0}\}$, where $\tilde{\mathbf{v}}_{\boldsymbol{\theta}i}$ denotes the ith row of $\tilde{\mathbf{V}}_{\boldsymbol{\theta}}$. Under some mild conditions, the BIC estimator (9) is consistent. Let $a_n = \max\{\theta_j, j \leq q\}$ and $b_n = \min\{\theta_j, j > q\}$, where the θ_j's are the penalty parameters defined in Section 1. The true model is indexed by s_T.

Theorem 1. *Let* $\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \mathrm{BIC}_{\boldsymbol{\theta}}$. *If Assumptions 1–3 in Appendix hold,* $\sqrt{n}a_n \xrightarrow{p} 0$ *and* $\sqrt{n}b_n \xrightarrow{p} \infty$, *then as* $n \to \infty$, $\Pr(s_{\widehat{\boldsymbol{\theta}}_n} = s_T) \to 1$.

The proof of this theorem is given in Appendix. The main idea of the proof is to compare the values between the considered model and the true model in two different cases according to whether the model is underfitted or overfitted. For penalized type estimators, Wang et al. [18], Wang and Leng [17] respectively established the consistency of SCAD and adaptive lasso estimators with the tuning parameter chosen by a BIC-type criterion. Unfortunately, their results were developed for the multiple regression model and thus are not directly applicable for (9) because the focus here is on subspaces rather than on coordinates. In practice, to avoid the p-dimensional tuning parameter selection [1], we usually let $\theta_i = \theta\|\widehat{\mathbf{v}}_i\|_2^{-r}$, where $\widehat{\mathbf{v}}_i$ is the ith row vector of the SSDRE $\widehat{\mathbf{V}}$ defined in (2), θ is a scalar tuning parameter, and $r > 0$ is some pre-specified parameter. As demonstrated in Appendix C, with this choice of tuning parameters, the result in Theorem 1 still holds. In the next section, this strategy is considered and $r = 0.5$ is used for illustration.

As a referee pointed, the problem (3) can be recovered as the MAP-solution based on the Multi-Laplacian prior [13]. Accordingly, it is possible to give a Bayesian treatment of the penalized SDR, allowing us to select the tuning parameters in a full Bayesian way (cf., [11]). This is beyond the scope of this paper but should definitely be a subject of future research.

3. Simulation study

In the simulation studies, we generated 500 datasets with different sample sizes n using models stated as follows:

Model 1

$$y = x_1/\{0.5 + (x_2 + 1.5)^2\} + \sigma\epsilon,$$

where $\epsilon \sim N(0, 1)$, $\mathbf{x} = (x_1, \ldots, x_{10})^T \sim N(0, \mathbf{\Sigma})$ with $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq 10$, and $\mathbf{x}$ and ϵ are independent. The scale parameter σ controls the signal-to-noise ratio and the parameter ρ is used to control the correlation among $\mathbf{x}$. In this model, the central subspace is spanned by the directions $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, \ldots, 0)^T$.

Model 2

$$\mathbf{x} = \mathbf{\Delta\Gamma}(y, y^2)^T + \tau\mathbf{\Delta}^{1/2}\boldsymbol{\epsilon},$$

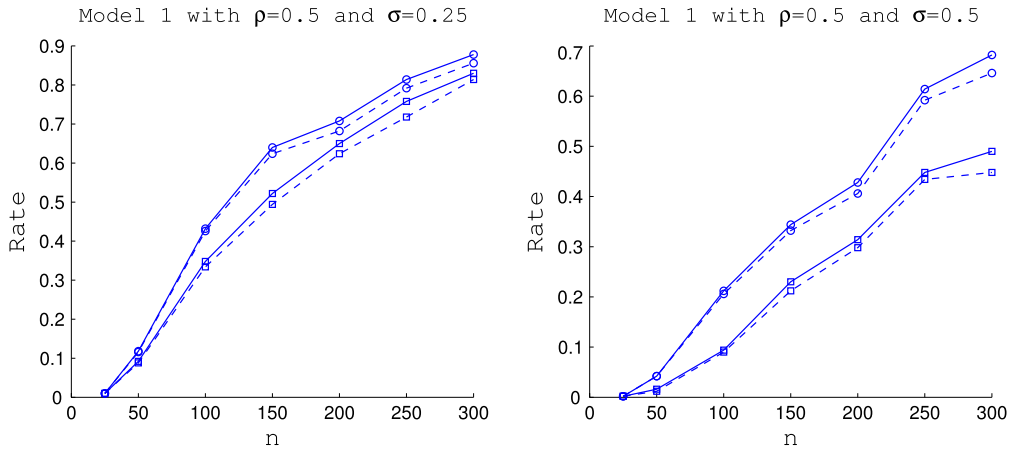

Fig. 1. The rate of selecting the true model versus n under Model 1 with $\rho = 0.5$ and $\sigma = 0.25$ (the left) or $\sigma = 0.5$ (the right). Solid and dashed lines represent the rates with BIC and AIC respectively. Lines marked with circles and squares indicate the solutions using PFC and SIR, respectively.

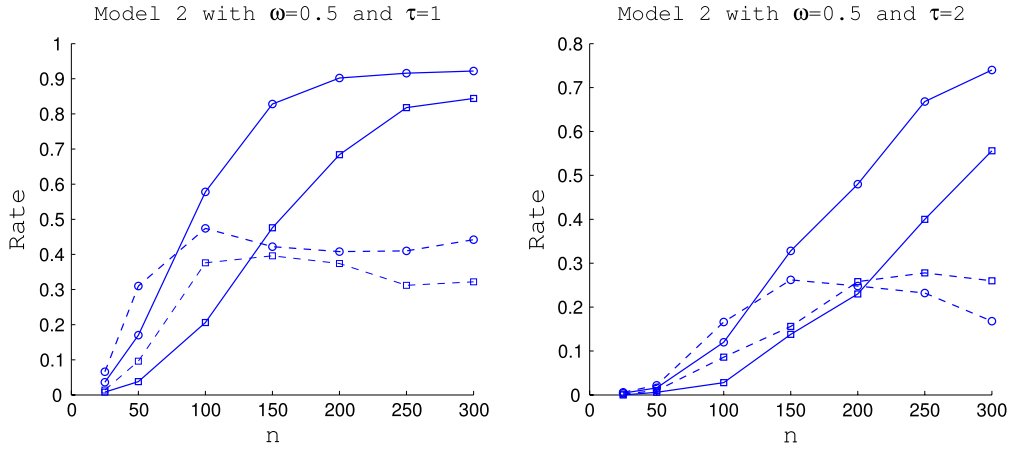

Fig. 2. The rate of selecting the true model versus n under Model 2 with $\omega = 0.5$ and $\tau = 1$ (the left) or $\tau = 2$ (the right). Solid and dashed lines represent the rates with BIC and AIC respectively. Lines marked with circles and squares indicate the solutions using PFC and SIR, respectively.

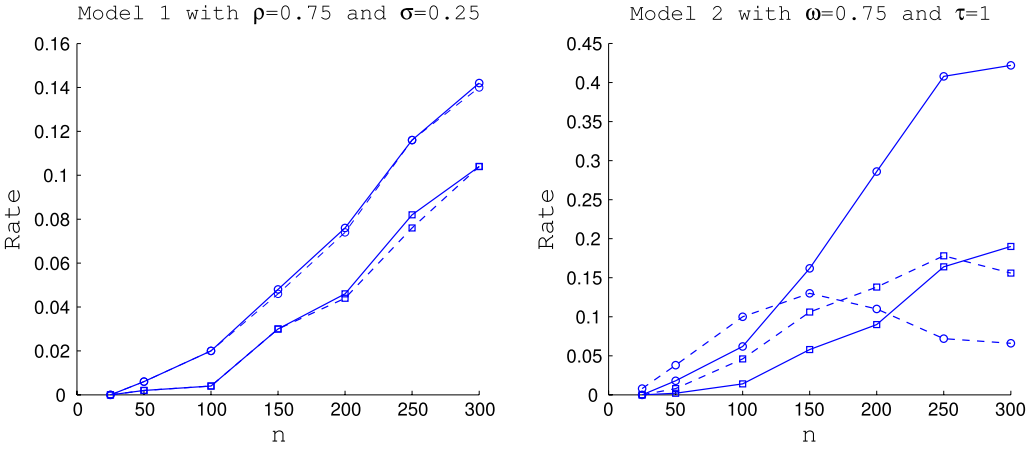

Fig. 3. The rate of selecting the true model versus n: Model 1 with $\rho = 0.75$ and $\sigma = 0.25$ (the left) and Model 2 with $\omega = 0.75$ and $\tau = 1$ (the right). Solid and dashed lines represent the rates with BIC and AIC respectively. Lines marked with circles and squares indicate the solutions using PFC and SIR, respectively.

where $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_{10}), y \sim N(0, 1), \Delta_{ij} = \omega^{|i-j|}$ for $1 \leq i, j \leq 10$, and y and $\boldsymbol{\epsilon}$ are independent. The first column of $\boldsymbol{\Gamma}$ is $(0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)^T$ and the second column of $\boldsymbol{\Gamma}$ is $(0.5, -0.5, 0.5, -0.5, 0, \ldots, 0)^T$. The scale parameter τ

controls the signal-to-noise ratio and the parameter ω is used to control the correlation among $\mathbf{x}$. In this model, the central subspace is the column space of $\mathbf{\Gamma}$.

We used SIR [6] and PFC [3] to generate $\mathbf{M}_n$ and $\mathbf{N}_n$ for CISE. Six slices were used for the SIR method. We calculated $\mathbf{M}_n$ in the PFC setting using fitted components $(|y|, y, y^2)^T$ for all simulation studies. Figs. 1–3 show the rate curves of selecting the true model (against the sample size n) using BIC under different model settings. For comparison, the corresponding rate curves with the AIC are also plotted in those figures. From Figs. 1–3, we can see the consistency of the CISE (PFC or SIR) with BIC in all settings, while the selection with AIC seems inconsistent for Model 2.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and two anonymous referees for their many helpful comments that have resulted in significant improvements in the article.

Appendix

Throughout this appendix, we use the following additional notations. Let $\mathbf{\Lambda}_{n1} = \mathrm{diag}\{\lambda_{n1}, \ldots, \lambda_{nd}\}$. The corresponding $\mathbf{\Lambda}_{n2}$ and the population analogs $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ can be understood. Since we use the matrix $\mathbf{V}$ to denote the basis of the subspace spanned by the columns of $\mathbf{V}$ in this paper, we use Chen et al.'s [1] definition of distance, $D(\mathbf{V}_n, \mathbf{V})$, which is defined as the largest eigenvalue of $(P_{\mathbf{V}_n} - P_{\mathbf{V}})^T(P_{\mathbf{V}_n} - P_{\mathbf{V}})$, where $P_{(\cdot)}$ represents the projection matrix with respect to the standard inner product.

Appendix A. Assumptions

Assumption 1. Let $\mathbf{V}_0$ denote the minimizer of (2) when the population matrices $\mathbf{M}$ and $\mathbf{N}$ are used instead. Then $\mathbf{V}_{0(s_T^c)} = 0$, where s_T^c is the complement of s_T in $\mathcal{S}$.

Assumption 2. $\mathbf{M}_n = \mathbf{M} + O_p(n^{-1/2})$ and $\mathbf{N}_n = \mathbf{N} + O_p(n^{-1/2})$.

Assumption 3. $\lambda_d > \lambda_{d+1}$.

These assumptions imposed here are all used for obtaining the oracle property of the CISE $\tilde{\mathbf{V}}$. They are mild and typically hold in many SDR methods. We refer to Chen et al. [1] for detailed discussions.

Appendix B. Lemmas

In order to prove Theorem 1, we firstly state two necessary lemmas.

Lemma 1. *Under Assumptions 1 and 2, for any $s \supset s_T$, $\mathbf{V}_{0(s)}$ is the minimizer of*

$$\arg\min_{\mathbf{V}\in\mathbb{R}^{p_s\times d}} -\mathrm{tr}(\mathbf{V}^T\mathbf{M}_{(s)}\mathbf{V}), \quad \textit{subject to } \mathbf{V}^T\mathbf{N}_{(s)}\mathbf{V} = \mathbf{I}_d. \tag{B.1}$$

Proof. Denote the solution of the minimization problem

$$\arg\min_{\{\mathbf{V}_s\in\mathbb{R}^{p\times d}:\mathbf{V}_s^T\mathbf{N}\mathbf{V}_s=\mathbf{I}_d\}} -\mathrm{tr}(\mathbf{V}_s^T\mathbf{M}\mathbf{V}_s) \tag{B.2}$$

as $\check{\mathbf{V}}$. Note that $\check{\mathbf{V}}_{(s^c)} = \mathbf{0}$ and $\check{\mathbf{V}}_{(s)}$ is the solution of (B.1) since $\check{\mathbf{V}}_{(s)}^T\mathbf{N}_{(s)}\check{\mathbf{V}}_{(s)} = \mathbf{I}_d$ and $\mathrm{tr}(\mathbf{V}_s^T\mathbf{M}\mathbf{V}_s) = \mathrm{tr}(\mathbf{V}_{s(s)}^T\mathbf{M}_{(s)}\mathbf{V}_{s(s)})$. On the other hand, because

$$\min_{\{\mathbf{V}\in\mathbb{R}^{p\times d}:\mathbf{V}^T\mathbf{N}\mathbf{V}=\mathbf{I}_d,\mathbf{V}_{(s^c)}=\mathbf{0}\}} -\mathrm{tr}(\mathbf{V}^T\mathbf{M}\mathbf{V}) \geq \min_{\{\mathbf{V}\in\mathbb{R}^{p\times d}:\mathbf{V}^T\mathbf{N}\mathbf{V}=\mathbf{I}_d\}} -\mathrm{tr}(\mathbf{V}^T\mathbf{M}\mathbf{V}), \tag{B.3}$$

and $\mathbf{V}_0$ satisfies the constraints in the left side of (B.3), we know $\mathbf{V}_0 = \check{\mathbf{V}}$. Then, the lemma immediately follows. □

Lemma 2. *Under Assumptions 1 and 2, for any $s \supset s_T$, we have $\lambda_i(s) = \lambda_i$ for $i = 1, \ldots, d$ and $\lambda_d(s) > \lambda_{d+1}(s)$, where $\lambda_1(s) \geq \cdots \geq \lambda_p(s) \geq 0$ are the eigenvalues of $\mathbf{G}(s)$.*

Proof. By the definition of the generalized eigenvalue problem, we have $\mathbf{M}\mathbf{V}_0 = \mathbf{N}\mathbf{V}_0\mathbf{\Lambda}_1$. Since $\mathbf{V}_{0(s^c)} = \mathbf{0}$, $\mathbf{M}_{(s)}\mathbf{V}_{0(s)} = \mathbf{N}_{(s)}\mathbf{V}_{0(s)}\mathbf{\Lambda}_1$. Also, by Lemma 1, $\mathbf{M}_{(s)}\mathbf{V}_{0(s)} = \mathbf{N}_{(s)}\mathbf{V}_{0(s)}\mathbf{\Lambda}_1(s)$, where $\mathbf{\Lambda}_1(s) = \mathrm{diag}\{\lambda_1(s), \ldots, \lambda_d(s)\}$. Thus, we have $\mathbf{\Lambda}_1(s) = \mathbf{\Lambda}_1$. Since $\mathbf{G}$ is a symmetric matrix, it follows from the well-known Sturm Theorem (Theorem 4.4.14; [9]) that $\lambda_{d+1}(s) \leq \lambda_{d+1} < \lambda_d$ by Assumption 3. □

Appendix C. Proof of Theorem 1

According to whether the resulting model $s_{\boldsymbol{\theta}}$ is underfitted, correctly fitted, or overfitted, we can partition $\mathbb{R}^{p+}$ into the following three mutually exclusive regions:

$$\mathbb{R}_U^{p+} = \{\boldsymbol{\theta} \in \mathbb{R}^{p+} : s_{\boldsymbol{\theta}} \not\supseteq s_T\},$$
$$\mathbb{R}_T^{p+} = \{\boldsymbol{\theta} \in \mathbb{R}^{p+} : s_{\boldsymbol{\theta}} = s_T\}, \quad \text{and}$$
$$\mathbb{R}_O^{p+} = \{\boldsymbol{\theta} \in \mathbb{R}^{p+} : s_{\boldsymbol{\theta}} \supset s_T, s_{\boldsymbol{\theta}} \neq s_T\}.$$

Moreover, for the purpose of proof, we could readily define a reference tuning parameter sequence $\boldsymbol{\theta}_* \in \mathbb{R}^{p+}$ which satisfies the conditions in Theorem 1. For instance, if we use Chen et al.'s [1] recommendation, say the adaptive-LASSO-type penalty, we could set the ith component of $\boldsymbol{\theta}_*$ as $\theta_{*i} = \beta_n \|\widehat{\mathbf{v}}_i\|_2^{-r}$ with a scalar sequence of tuning parameters β_n satisfying $\sqrt{n}\beta_n \to 0$ and $n^{(1+r)/2}\beta_n \to \infty$, where $r > 0$ is some pre-specified parameter. By Theorem 2-(i) in [1], $s_{\boldsymbol{\theta}_*} \in \mathbb{R}_T^{p+}$ with probability tending to 1. Thus, to prove the theorem, it suffices to show that $\Pr(\inf_{\boldsymbol{\theta} \in \mathbb{R}_U^{p+} \cup \mathbb{R}_O^{p+}} \mathrm{BIC}_{\boldsymbol{\theta}} > \mathrm{BIC}_{\boldsymbol{\theta}^*}) \to 1$. The following proof consists of two steps.

Step 1. Let us firstly consider $\boldsymbol{\theta} \in \mathbb{R}_O^{p+}$. We then have $p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}_*} \geq 1$. We shall show that with probability approaching one, the BIC favors $s_{\boldsymbol{\theta}_*}$. Before proceeding, to facilitate our proof we need another definition, the unpenalized SDRE under the model identified by $\tilde{\mathbf{V}}_{\boldsymbol{\theta}}$, say

$$\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}} = \underset{\{\mathbf{V} \in \mathbb{R}^{p \times d} : \mathbf{V}_{(s_{\boldsymbol{\theta}}^c)} = \mathbf{0}\}}{\arg\min} -\mathrm{tr}(\mathbf{V}^T \mathbf{M}_n \mathbf{V}), \quad \text{subject to } \mathbf{V}^T \mathbf{N}_n \mathbf{V} = \mathbf{I}_d.$$

By this definition, we must have $g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}}; \mathbf{M}_n) \geq g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n)$. Thus, write

$$\begin{aligned} n(\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*}) &= n[g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n)] + d(p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}_*}) \cdot \log n \\ &\geq n[g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n)] + d(p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}_*}) \cdot \log n \\ &\geq n[g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n)] + \log n. \end{aligned} \tag{C.4}$$

Next, we will show $n[g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n)] = O_p(1)$. By Theorem 1 in [1], we know that $n[g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n) - g(\mathbf{V}_0; \mathbf{M}_n)] = O_p(1)$. It remains to examine $n[g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\mathbf{V}_0; \mathbf{M}_n)]$.

By using similar arguments in Section 2.2, finding $\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}$ is equivalent to the minimization problem

$$\widehat{\boldsymbol{\Gamma}}_{s_{\boldsymbol{\theta}}} = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{p s_{\boldsymbol{\theta}} \times d}}{\arg\min} -\mathrm{tr}(\boldsymbol{\Gamma}^T \mathbf{G}_n(s_{\boldsymbol{\theta}}) \boldsymbol{\Gamma}), \quad \text{subject to } \boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_d.$$

Then $\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}(s_{\boldsymbol{\theta}})} = \mathbf{N}_{n(s_{\boldsymbol{\theta}})}^{-1/2} \widehat{\boldsymbol{\Gamma}}_{s_{\boldsymbol{\theta}}}$ and $\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}(s_{\boldsymbol{\theta}}^c)} = \mathbf{0}$. Denote $\boldsymbol{\Gamma}_*$ as an orthonormal basis matrix of the subspace spanned by the columns of $\mathbf{N}_{n(s_{\boldsymbol{\theta}})}^{1/2} \mathbf{V}_{0(s_{\boldsymbol{\theta}})}$. Thus, there exists a positive definite matrix $\mathbf{O} \in \mathbb{R}^{d \times d}$ so that $\boldsymbol{\Gamma}_* = \mathbf{N}_{n(s_{\boldsymbol{\theta}})}^{1/2} \mathbf{V}_{0(s_{\boldsymbol{\theta}})} \mathbf{O}$. It suffices to prove $D(\widehat{\boldsymbol{\Gamma}}_{s_{\boldsymbol{\theta}}}, \boldsymbol{\Gamma}_*) = O_p(n^{-1/2})$. By Assumption 1, we have $\mathbf{V}_{0(s_{\boldsymbol{\theta}})}^T \mathbf{N}_{(s_{\boldsymbol{\theta}})} \mathbf{V}_{0(s_{\boldsymbol{\theta}})} = \mathbf{I}_d$ since $s_T \subset s_{\boldsymbol{\theta}}$. In addition, by Assumption 2, $\mathbf{O}^T \mathbf{O} = \mathbf{I}_d + O_p(n^{-1/2})$.

Similar to the techniques in the proof of Theorem 1 in [1], it suffices to show, for any arbitrarily small $\varepsilon > 0$, there exits a sufficiently large constant C, such that

$$\liminf_n \Pr\left(\inf_{\mathbf{Z} \in T_{\boldsymbol{\Gamma}_*}(p_{s_{\boldsymbol{\theta}}}, d) : \|\mathbf{B}\|_t = C} g(R(\boldsymbol{\Gamma}_* + n^{-1/2}\mathbf{Z}); \mathbf{G}_n(s_{\boldsymbol{\theta}})) > g(\boldsymbol{\Gamma}_*; \mathbf{G}_n(s_{\boldsymbol{\theta}})) \right) > 1 - \varepsilon. \tag{C.5}$$

By using Lemma 1-(ii) in [1] again, for $\mathbf{Z} \in T_{\boldsymbol{\Gamma}_*}(p_{s_{\boldsymbol{\theta}}}, d)$ we have

$$\begin{aligned} &n\left\{ g(R(\boldsymbol{\Gamma}_* + n^{-1/2}\mathbf{Z}); \mathbf{G}_n(s_{\boldsymbol{\theta}})) - g(\boldsymbol{\Gamma}_*; \mathbf{G}_n(s_{\boldsymbol{\theta}})) \right\} \\ &\quad = \left[-\mathrm{tr}(\mathbf{Z}^T \mathbf{G}_n(s_{\boldsymbol{\theta}}) \mathbf{Z}) - 2\sqrt{n}\mathrm{tr}(\mathbf{Z}^T \mathbf{G}_n(s_{\boldsymbol{\theta}}) \boldsymbol{\Gamma}_*) + \mathrm{tr}(\mathbf{Z}^T \mathbf{Z} \boldsymbol{\Gamma}_*^T \mathbf{G}_n(s_{\boldsymbol{\theta}}) \boldsymbol{\Gamma}_*) \right] (1 + o_p(1)) \\ &\quad := (\Delta_1 + \Delta_2 + \Delta_3)(1 + o_p(1)). \end{aligned}$$

Denote $\boldsymbol{\Gamma}_0(s_{\boldsymbol{\theta}}) = \mathbf{N}_{(s_{\boldsymbol{\theta}})}^{1/2} \mathbf{V}_{0(s_{\boldsymbol{\theta}})}$. Based on Lemma 1 and Assumption 2, using similar arguments as in the proof of Theorem 1 in [1], it can be verified that

$$\begin{aligned} \Delta_1 + \Delta_3 &\geq [\lambda_d(s_{\boldsymbol{\theta}}) - \lambda_{d+1}(s_{\boldsymbol{\theta}})] \|\mathbf{B}\|_t^2 + o_p(1), \\ \Delta_2 &= \sqrt{n}\mathrm{tr}\{\mathbf{B}^T \boldsymbol{\Gamma}_{0\perp}^T(s_{\boldsymbol{\theta}}) [\mathbf{G}_n(s_{\boldsymbol{\theta}}) - \mathbf{G}(s_{\boldsymbol{\theta}})] \boldsymbol{\Gamma}_0(s_{\boldsymbol{\theta}})\} (1 + O_p(n^{-1/2})). \end{aligned}$$

As a consequence, by the Cauchy–Schwarz inequality for a trace operator, Δ_2 is uniformly bounded by $\|\mathbf{B}\|_t \times \|\sqrt{n}[\mathbf{G}_n(s_{\boldsymbol{\theta}}) - \mathbf{G}(s_{\boldsymbol{\theta}})] \boldsymbol{\Gamma}_0(s_{\boldsymbol{\theta}})\|_t$. Therefore, as long as the constant C is sufficiently large, $\Delta_1 + \Delta_3$ will always dominate Δ_2 with arbitrarily

large probabilities by Lemma 2. This implies the inequality (C.5), and as a consequence we will have $n[g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n)]$ is of order $O(1)$. As a result, $\Pr(\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*} > 0) \to 1$ for any $\boldsymbol{\theta} \in \mathbb{R}_O^{p+}$ since the last term in (C.4) diverges to infinity as $n \to \infty$.

Step 2. Now consider $\boldsymbol{\theta} \in \mathbb{R}_U^{p+}$. In this case, there at least exists $i \in s_T$ so that $\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}(i)} = \mathbf{0}$. Thus, $D(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}, \mathbf{V}_0) = C$ for some constant C. Consequently, $D(\check{\mathbf{V}}, \mathbf{V}_0) = C + O_p(n^{-1/2})$, where $\check{\mathbf{V}}$ is defined in (B.2). Similar to Step 1, we have

$$\begin{aligned}
\mathrm{BIC}_{\boldsymbol{\theta}} - \mathrm{BIC}_{\boldsymbol{\theta}_*} &\geq g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}_n) - g(\tilde{\mathbf{V}}_{\boldsymbol{\theta}_*}; \mathbf{M}_n) - dp_{\boldsymbol{\theta}_*} \cdot \log n/n \\
&= g(\widehat{\mathbf{V}}_{s_{\boldsymbol{\theta}}}; \mathbf{M}) - g(\mathbf{V}_0; \mathbf{M}) - dp_{\boldsymbol{\theta}_*} \cdot \log n/n + O_p(n^{-1/2}) \\
&\xrightarrow{p} g(\check{\mathbf{V}}; \mathbf{M}) - g(\mathbf{V}_0; \mathbf{M}) > 0,
\end{aligned}$$

in which the last inequality comes from the fact that $\mathbf{V}_0$ is the minimizer of $\min_{\mathbf{V}} -\mathrm{tr}(\mathbf{V}^T\mathbf{G}\mathbf{V})$ subject to $\mathbf{V}^T\mathbf{N}\mathbf{V} = \mathbf{I}_d$ by definition. Thus, it follows immediately that $\Pr(\mathrm{BIC}_{\boldsymbol{\theta}} > \mathrm{BIC}_{\boldsymbol{\theta}_*}) \to 1$ for any $\boldsymbol{\theta} \in \mathbb{R}_U^{p+}$.

Combining the two cases together implies that any $\boldsymbol{\theta}$ failing to identify the true model cannot be selected as the optimal parameter. Say, the model associated with the optimal $\boldsymbol{\theta}$ must be the true one which completes the proof. □

References

[1] X. Chen, C. Zou, R.D. Cook, Coordinate-independent sparse sufficient dimension reduction and variable selection, Ann. Statist. 38 (2010) 3696–3723.
[2] R.D. Cook, Regression Graphics: Ideas for Studying Regressions Through Graphics, Wiley, New York, 1998.
[3] R.D. Cook, L. Forzani, Principal fitted components for dimension reduction in regression, Statist. Sci. 23 (2008) 485–501.
[4] A. Edelman, T.A. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Math. Anal. 20 (1998) 303–353.
[5] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[6] K.-C. Li, Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. 86 (1991) 316–327.
[7] L. Li, Sparse sufficient dimension reduction, Biometrika 94 (2007) 603–613.
[8] J.H. Manton, Optimization algorithms exploiting unitary constraints, IEEE Trans. Signal Process. 50 (2002) 635–650.
[9] M. Marcus, H. Minc, Survey of Matrix Theory and Matrix Inequalities, Dover, New York, 1992.
[10] L. Ni, R.D. Cook, C.L. Tsai, A note on shrinkage sliced inverse regression, Biometrika 92 (2005) 242–247.
[11] T. Park, G. Casella, The Bayesian lasso, J. Amer. Statist. Assoc. 103 (2008) 681–686.
[12] A.E. Raftery, Approximate Bayes factors and accounting for model uncertainty in generalized linear models, Biometrika 83 (1996) 251–266.
[13] S. Raman, T.J. Fuchs, P.J. Wild, E. Dahl, V. Roth, The Bayesian group-lasso for analyzing contingency tables, in: ICML'09 Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 881–888.
[14] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[15] L. Tierney, J.B. Kadane, Accurate approximations for posterior moments and marginal densities, J. Amer. Statist. Assoc. 81 (1986) 82–86.
[16] L. Wang, Wilcoxon-type generalized Bayesian information criterion, Biometrika 96 (2009) 163–173.
[17] H. Wang, C. Leng, Unified lasso estimation by least square approximation, J. Amer. Statist. Assoc. 102 (2007) 1039–1048.
[18] H. Wang, R. Li, C.L. Tsai, On the consistency of SCAD tuning parameter selector, Biometrika 94 (2007) 553–568.
[19] L. Zhu, L. Zhu, Nonconcave penalized inverse regression in single-index models with high dimensional predictors, J. Multivariate Anal. 100 (2009) 862–875.