Optimal variable selection in regression models

Jie Ding

School of Statistics, University of Minnesota, USA

E-mail: dingj@umn.edu

Vahid Tarokh

Department of Electrical and Computer Engineering, Duke University, USA

E-mail: vahid.tarokh@duke.edu

Yuhong Yang

School of Statistics, University of Minnesota, USA

E-mail: yangx374@umn.edu

Summary. We introduce a new criterion for variable selection in regression models, and show its optimality in terms of both loss and risk under appropriate assumptions. The key idea is to impose a penalty that is nonlinear in model dimensions. In contrast to the state-of-art model selection criteria such as the C_p method, delete-1 or delete-k cross-validation, Akaike information criterion, Bayesian information criterion, the proposed method is able to achieve asymptotic loss and risk efficiency in both parametric and nonparametric regression settings, giving new insights on the reconciliation of two types of classical criteria with different asymptotic behaviors. Adaptivity and wide applicability of the new criterion are demonstrated by several numerical experiments. Unless the signal to noise ratio is very low, it performs better than some popular methods in our experimental study. An R package 'bc' is released that serves as a supplement to this work.

Keywords: Asymptotic efficiency, Consistency, Regression, Subset selection, Variable selection

1. Introduction

Consider the task of estimating the regression function $f(\boldsymbol{x}) = E(Y \mid X = \boldsymbol{x})$. With a given class of linear models, indexed by $\alpha \in \mathcal{A}_n$ where n is the number of observations, we need to select one of them to best capture the underlying distribution of the data or best predict the future response. Suppose that each α is a subset of $\{1, \ldots, d_n\}$, and that the least squares estimator (LSE) is used to fit each candidate model. Given the design matrix $X = [\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}}]^{\mathrm{T}}$ and observations $\boldsymbol{y}_n = [y_1, \ldots, y_n]^{\mathrm{T}}$, our goal is to select $\alpha \in \mathcal{A}_n$ that minimizes the squared error loss $\mathcal{L}_n(\alpha) = n^{-1}\|\boldsymbol{f}_n - \hat{\boldsymbol{f}}_n(\alpha)\|^2$, or the risk $\mathcal{R}_n(\alpha) = E\{\mathcal{L}_n(\alpha)\}$ (where the expectation is with respect to random noises) as much as possible. Here, $\boldsymbol{f}_n \overset{\Delta}{=} [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^{\mathrm{T}}$, $\hat{\boldsymbol{f}}_n(\alpha)$ is the LSE of $\boldsymbol{f}_n$ in model α, and $\|\cdot\|$ denotes the Euclidean norm. The above framework includes the usual variable selection and subset selection in linear regression, and the selection of basis such as polynomials, splines, or wavelets in function estimation. We also note that approaching the minimal loss or risk is usually equivalent to achieving consistency in variable selection when the true data generating model is inside $\mathcal{A}_n$, as elaborated later in the paper.

A wide variety of criteria for variable selection have been proposed in the literature, motivated from different viewpoints and justified under various circumstances. Comparisons of merits and shortcomings of these methods have flourished in the past decades (Stone, 1979; Shibata, 1981; Nishii et al., 1984; Li, 1987; Rao and Wu, 1989; Speed and Yu, 1993; Shao, 1993; Yang and Barron, 1998). A detailed summary can be found in the work of Shao (1997). These methods typically fall into two classes according to their asymptotic performances. Methods in the first class achieve consistency, in the sense that the correct model with the smallest dimension is selected with probability going to one as n tends to infinity. However, they usually perform sub-optimally when $\mathcal{A}_n$ does not contain any data generating model (a correct model). Examples include the Bayesian information criterion (BIC) (Schwarz, 1978), minimum description length (MDL) criterion (Barron et al., 1998; Hansen and Yu, 2001), Bayes factors (Casella et al., 2009), the delete-k cross-validation (CV) method with $\lim_{n\rightarrow\infty} k/n = 1$ (Geisser, 1975; Burman, 1989; Shao, 1993; Zhang, 1993), Generalized information criterion (GIC$_{\lambda_n}$) with $\lambda_n \rightarrow \infty$ (Nishii et al., 1984; Rao and Wu, 1989). Some other methods motivated from the literature of autoregressive order selection include the Hannan and Quinn (HQ) criterion (Hannan and Quinn, 1979), the predictive minimum description length (PMDL) criterion (Rissanen, 1986; Wei, 1992), and the predictive least squares (PLS) principle (Wei, 1992).

Methods in the second class usually achieve asymptotic efficiency, in the sense that their predictive performance are asymptotically equivalent to the best offered by the candidate models, when $\mathcal{A}_n$ contains no more than one correct model. They tend to overfit when there are at least two correct candidate models. In other words, the smallest correct model cannot be selected with probability going to one as n increases. Examples include the Akaike information criterion (AIC) (Akaike, 1970, 1998), C_p method (Mallows, 1973), final prediction error (FPE) criterion (Akaike, 1969), the generalized CV (GCV) method (Craven and Wahba, 1978), the delete-1 CV method (Stone, 1977) (or leave-one-out, LOO), and GIC$_{\lambda_n}$ with $\lambda_n = 2$ (Shao, 1997). One can define another class by considering, for example, delete-k CV with $k/n \rightarrow \rho \in (0, 1)$ and GIC$_{\lambda_n}$ with $\lambda_n \neq 2$ being a constant. But these criteria usually do not exhibit asymptotic efficiency in typical situations of interest. From the above, the following question naturally arises:

Is it even possible to adaptively achieve the better performance of the two classes in all situations? Also, for a variety of choices of $\mathcal{A}_n$, which may or may not contain correct model(s), is there a simple way to consitently select the model that attains the lower bound of $\mathcal{L}_n(\alpha)$ or $\mathcal{R}_n(\alpha)$ as n tends to infinity? These questions are important because in real applications, usually a data analyst does not know whether the data generating model is correctly specified or even finite dimensional.

As was discussed before, the ability to consistently identify the smallest correct model when $\mathcal{A}_n$ contains at least two correct models is the typical watershed of the two classes. We note that consistency in selection implies asymptotic loss/risk efficiency (which will be elaborated in the next section). Thus, an ideal model selection criterion that combines the merits of both classes would behave in the following manner. It achieves consistency whenever $\mathcal{A}_n$ contains at least one correct model (for all sufficiently large n), and asymptotic efficiency whenever $\mathcal{A}_n$ does not contain any correct model. The above two situations are also referred to as "parametric" and "nonparametric", respectively. Throughout the paper, we allow the data generating models (and their dimensions in the parametric case) to be dependent on n. Fortunately, there have been some work towards the direction of adaptive selection procedures. One approach is to take data-dependent penalties that bring in adaptation capabilities (Barron et al., 1994; Hansen and Yu, 2001). Yang (2007) proposed an approach that examines whether BIC selects the same model successively at different sample sizes, in order to adaptively achieve asymptotic efficiency in both parametric and nonparametric situations. Ing (2007) proposed a hybrid selection procedure combining AIC and BIC in the context of order selection of autoregressive models. A measure called parametricness index was proposed by Liu and Yang (2011) to adaptively switch between AIC and BIC regimes. Erven et al. (2012) proposed to use a switching distribution to perform sequential Bayesian model averaging or to encourage early switch to a better model. Their predictive approach can adaptively achieve the optimal cumulative risk convergence rates and thus provide a Bayesian remedy for the AIC-BIC dilemma. Zhang and Yang (2015) proposed a CV procedure for choosing between model selection criteria, and showed that the hybrid criterion asymptotically behaves like the better one of AIC and BIC under a suitably chosen data splitting ratio. In a recent work, Ding et al. (2018) proposed a criterion referred to as Bridge criterion for autoregressive order selection. The main idea is to penalize different model dimensions with nonlinear penalty terms, in contrast to the linear terms used by AIC and BIC. In this work, we introduce a new criterion for regression variable selection or subset selection, motivated by a similar idea in (Ding et al., 2018). We stress, however, that the results developed for autoregressive order selection problem can not be trivially applied to regression problems. The candidate models as represented by subsets of variables are usually non-nested, and the form of the criterion is going to be different (as much different as C_p is from AIC). We will also extend the proposed approach to high dimensional regression. But due to a similar spirit in choosing the penalty terms, we shall also refer to the method as Bridge criterion (BC).

The purpose of this work is to provide a theoretical possibility that the two classes of model selection criteria can be reconciled in one criterion for regression problems. As was summarized by Shao (1997), the classical criteria may be written in a form that involves a penalty term proportional to the dimension of each model. In contrast, a key element of the introduced BC criterion is that the penalty term is proportional to $1 + 2^{-1} + \cdots + d^{-1}$ for each model of dimension d. As we shall see, employment of this harmonic number makes it intrinsically different with any existing model selection criterion, and also "bridges" the features of the two classes. Under some regularity conditions, we show that BC achieves the consistency in parametric settings and asymptotic efficiency in nonparametric settings. As a result, it achieves the asymptotic loss and risk efficiency in rather general situations.

The outline of this paper is given below. In Section 2, we propose the new variable selection criterion and define a measure called parametricness index, along with intuitive explanations. In Section 3, we review the GIC_{λ_n} method with $\lambda_n = 2$ and $\lambda_n \to \infty$, with some extensions of existing results. They serve as the two representatives of various state-of-art methods. And based on those extensions, we rigorously prove the asymptotic loss and risk optimality of the new criterion in various settings under reasonable assumptions. In Section 4, an adaptation of the proposed method is further applied to high dimensional variable selection where sample size could be smaller than model dimensions. Numerical results are given in Section B demonstrating the performances of the proposed method. We make our conclusions in Section 5, and outline some discussions on future work.

2. A new variable selection criterion

We define

$$S_n(\alpha) = \|\boldsymbol{y}_n - \hat{\boldsymbol{f}}_n(\alpha)\|^2, \tag{1}$$

and let $\hat{\sigma}_n^2$ be an estimator of σ^2. Many model selection procedures are equivalent or closely related to the following *Generalized information criterion* (or GIC_{λ_n} procedure) which selects

$$\hat{\alpha}_n = \arg\min_{\alpha \in \mathcal{A}_n} G_{n,\lambda_n}(\alpha) \triangleq \frac{S_n(\alpha)}{n} + \frac{\lambda_n \hat{\sigma}_n^2 d_n(\alpha)}{n}, \tag{2}$$

where λ_n is a deterministic sequence of n that controls the trade-off between the goodness-of-fit and model parsimoniousness (Shao, 1997). If there exist more than one minimizers in (2), we arbitrarily choose one of them as $\hat{\alpha}_n$ (e.g. the one with the smallest dimension). Recall that $\bar{\alpha}_n = \{1, \ldots, d_n\}$. With $\hat{\sigma}_n^2 \triangleq (n - d_n)^{-1} S(\bar{\alpha}_n)$, GIC$_2$ reduces to the C_p method (Mallows, 1973) and GIC$_{\lambda_n}$ with $n^{-1}\lambda_n + \lambda_n^{-1}(\log\log n) \to 0$ is the GIC method proposed by Rao and Wu (1989). If we replace $\hat{\sigma}_n^2$ with $\{n - d_n(\alpha)\}^{-1} S(\alpha)$ and assume $n^{-1}\lambda_n d_n + n^{-1/2} d_n \to 0$, then a direct calculation (using $\log(1+x) = x + o(x)$) shows that (2) is basically equivalent to minimizing

$$\log \frac{S_n(\alpha)}{n} + \frac{\lambda_n d_n(\alpha)}{n}. \tag{3}$$

In this case, $\lambda_n = 2$ corresponds to AIC and $\lambda_n = \log n$ corresponds to BIC. Moreover, AIC was shown to be asymptotically equivalent to delete-1 CV (Stone, 1977) and GCV if $d_n = o(n)$ (Shao, 1997). In general, delete-k CV has the same asymptotic behavior as the GIC$_{\lambda_n}$ with (Shao, 1997)

$$\lambda_n = \frac{n}{n-k} + 1. \tag{4}$$

We propose the following Bridge criterion (BC):

$$\hat{\alpha}_{\mathrm{BC}} = \underset{\alpha \in \mathcal{A}_n, d_n(\alpha) \leq d_n(\hat{\alpha}_{\mathrm{GIC}_2})}{\arg\min} B_{n,\lambda_n}(\alpha) \triangleq \frac{S_n(\alpha)}{n} + \frac{\lambda_n \hat{\sigma}_n^2 H_{d_n(\alpha)}}{n} \tag{5}$$

where $\hat{\alpha}_{\mathrm{GIC}_2}$ is the model selected by GIC$_2$ procedure, and $H_{d_n(\alpha)} = \sum_{k=1}^{d_n(\alpha)} k^{-1}$. The major difference with GIC$_{\lambda_n}$ and many others is that the penalty function of BC is nonlinear in model dimension. More details on $\hat{\sigma}_n^2$, λ_n, and the performance of BC and GIC will be examined in the next section.

Below we provide some intuitive explanations. In BC, we start by imposing a GIC$_{\lambda_n}$-type ($\lambda_n \to \infty$) heavy penalty, but alleviate it more and more to endow the selection procedure with the following *self-awareness.*

1) if the model class is eventually parametric, some parsimonious model ($\underline{\alpha}_n^c$) is already adequately explaining the data, so that extra dimensions become more and more obviously redundant; In other words, the models likely to be selected by BC are of dimensions no more than $O_p(1)$ above $d_n(\underline{\alpha}_n^c)$. But, within that region, the dispensable candidates suffer from extra penalties

$$\frac{\lambda_n \hat{\sigma}_n^2}{n}\left(\frac{1}{d_n(\underline{\alpha}_n^c)+1}+\cdots+\frac{1}{d_n(\alpha)}\right)=\frac{\lambda_n}{d_n(\underline{\alpha}_n^c)}\frac{\hat{\sigma}_n^2}{n}O_p(1) \tag{6}$$

compared with (the optimal one) $\underline{\alpha}_n^c$. As long as $\lambda_n/d_n(\underline{\alpha}_n^c) \to \infty$ and $\hat{\sigma}_n \not\to_p 0$, (6) resembles that of GIC_{λ_n}, thus exhibiting similar asymptotic behavior.

2) if the model class is nonparametric, any model of a fixed dimension is usually not able to take advantage of more and more observations; thus, increasingly higher dimensions are likely to be selected. In other words, larger models tend to be favored, and this is even more accelerated by the smaller and smaller penalty increments induced by $H_{d_n(\alpha)}$. As a result, if λ_n diverges not too rapidly, BC selects the largest model $\hat{\alpha}_{\mathrm{GIC}_2}$, which is asymptotically loss and risk efficient.

The theoretical analysis of BC will be elaborated in the next section. We shall show that for a wide variety of λ_n's and parametric/nonparametric situations, BC enjoys the universal optimality (in terms of handling both parametric and nonparametric settings). We suggest $\lambda_n = n^{1/3}$ based on our extensive numerical experiments. A reasonable baseline level of performance is achieved without further tuning of λ_n.

Building upon the proposed criterion, we define the following quantity referred to as parametricness index (PI). $\mathrm{PI}_n = 1$ if $d_n(\hat{\alpha}_{\mathrm{GIC}_2}) = d_n(\hat{\alpha}_{\mathrm{GIC}_{\lambda_n}}) = d_n(\hat{\alpha}_{\mathrm{BC}})$, and

$$\mathrm{PI}_n = \frac{d_n(\hat{\alpha}_{\mathrm{GIC}_2}) - d_n(\hat{\alpha}_{\mathrm{BC}})}{d_n(\hat{\alpha}_{\mathrm{GIC}_2}) - d_n(\hat{\alpha}_{\mathrm{BC}}) + |d_n(\hat{\alpha}_{\mathrm{BC}}) - d_n(\hat{\alpha}_{\mathrm{GIC}_{\lambda_n}})|}$$

otherwise, given a prescribed candidate set $\mathcal{A}_n$. The λ_n in the above GIC_{λ_n} is chosen to be a diverging sequence to ensure consistency in the parametric scenario. From the literature, we suggest $\lambda_n = \log n$ so that GIC_{λ_n} performs closely to BIC. Following our definition, $\textsc{pi}_n \in [0, 1]$. Intuitively, $\textsc{pi}_n$ is close to one in parametric scenario where $d_n(\hat{\alpha}_{\textsc{bc}}), d_n(\hat{\alpha}_{\textsc{gic}_{\lambda_n}})$ do not differ much, while close to zero in nonparametric scenario where $d_n(\hat{\alpha}_{\textsc{bc}})$ and $d_n(\hat{\alpha}_{\textsc{gic}_2})$ are close and larger than $d_n(\hat{\alpha}_{\textsc{gic}_{\lambda_n}})$.

The goal of PI is to measure the extent to which the specified model class is adequate in explaining the observed data, namely to assess the confidence that the selected model can be practically treated as the data-generating model (given sample size n). The larger $\textsc{pi}_n$, the more confidence. In parametric situations where BC and GIC_{λ_n} are consistent in variable selection, clearly we have $\textsc{pi}_n \rightarrow_p 1$. Also, whether $\mathbb{P}\{\textsc{pi}_n \leq t\} \rightarrow 1$ for some constant $t < 1$ in nonparametric situations depends on more knowledge about $\mathcal{R}_n(\cdot)$. We shall provide Proposition 3 in Section 3 that indicates the following simple rule for answering the question if we are in a parametric or nonparametric scenario: regard it parametric if $\textsc{pi}_n > t$ for some $0 < t < 1$ and nonparametric otherwise.

3. Asymptotic performance of classical criteria and BC

In this section, we review the asymptotic performance of classical criteria in parametric and nonparametric situations, and provide some extensions of existing results. In addition, we show that no existing method, or any new one for that matter, can be consistent in selection of the best model for nonparametric situations. Based on this, for adaptation over parametric and nonparametric situations, while adaptive selection consistency is clearly ruled out, there is still hope to achieve asymptotic efficiency adaptively. We then prove that the proposed Bridge criterion can achieve such goal.

3.1. Regression models and goals

Let $\mathcal{A}_n$ denote the class of candidate models. Without loss of generality, we assume that $\bar{\alpha}_n = \{1, \ldots, d_n\}$ is the largest model in $\mathcal{A}_n$. It is common to allow d_n to diverge with n in order to include any correct model with fixed dimension (if there exists), and to achieve optimal loss/risk when no candidate model is correct. We assume that $X_n^{\mathrm{T}} X_n$, $n = 1, 2, \ldots$ are non-random and invertible. We consider the classical setting $n > d_n$ for now, until we generalize it in Section 4. We refer to $d_n(\alpha)$ as the dimension of model α. Each $\alpha \in \mathcal{A}_n$ denotes the candidate linear model that assumes $\boldsymbol{f}_n$ is in the column linear span of $X_n(\alpha)$, the corresponding $n \times d_n(\alpha)$ sub-matrix of X_n. In other words, the LSE of $\boldsymbol{f}_n$ under model α is $\hat{\boldsymbol{f}}_n(\alpha) = P_n(\alpha)\boldsymbol{y}_n$, where $P_n(\alpha) = X_n(\alpha)\{X_n(\alpha)^{\mathrm{T}} X_n(\alpha)\}^{-1} X_n(\alpha)^{\mathrm{T}}$. We define

$$\Delta_n(\alpha) = n^{-1} \| P_n(\alpha)^{\perp} \boldsymbol{f}_n \|^2 \tag{7}$$

and refer to it as the model approximation error. A candidate model α is correct if it satisfies $\Delta_n(\alpha) = 0$. Let $\mathcal{A}_n^c$ denote the set of all the correct candidate models in $\mathcal{A}_n$. Let $\boldsymbol{e}_n = \boldsymbol{y}_n - \boldsymbol{f}_n = [e_1, \ldots, e_n]^{\mathrm{T}}$ denote the noise vector. We assume that $e_1, \ldots, e_n$ are independent and identically distributed (i.i.d.) with variance $\sigma^2 > 0$.

Throughout this paper, the model class is referred to as "parametric" if $\mathcal{A}_n^c \neq \emptyset$, and "nonparametric" if $\mathcal{A}_n^c = \emptyset$, for all sufficiently large n. Subsection B.1 includes a specific example showing the parametric and nonparametric settings. If $\mathcal{A}_n^c \neq \emptyset$, we let $\underline{\alpha}_n^c$ denote the model in $\mathcal{A}_n^c$ with the smallest dimension. We make the following assumptions. For all sufficiently large n,

$$\underline{\alpha}_n^c \subseteq \alpha, \forall \alpha \in \mathcal{A}_n^c \setminus \{\underline{\alpha}_n^c\} \quad \text{if } \mathcal{A}_n^c \neq \{\underline{\alpha}_n^c\}, \tag{8}$$

$$\max_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \frac{\mathcal{R}_n(\underline{\alpha}_n^c)}{\mathcal{R}_n(\alpha)} \to 0 \quad \text{if } 0 < \text{card}(\mathcal{A}_n^c) < \text{card}(\mathcal{A}_n). \tag{9}$$

Assumption (8) guarantees that $\underline{\alpha}_n^c$ is uniquely defined, and that $\underline{\alpha}_n^c$ minimizes $\mathcal{R}_n(\alpha)$ over $\mathcal{A}_n^c$. Assumption (9) is required to distinguish the "parametric" from the "nonparametric". Instead of assumptions (8) and (9), we may also assume that $\underline{\alpha}_n^c$ is fixed and there exists a fixed vector $\boldsymbol{\beta}$ (which does not depend on n) satisfying

$$\boldsymbol{f}_n = X_n(\underline{\alpha}_n^c)\boldsymbol{\beta}, \text{ eig}_{\min}(X_n^{\mathrm{T}} X_n) \sim n, \text{ eig}_{\max}(X_n^{\mathrm{T}} X_n) \sim n, \; n^{-1} d_n \to 0. \tag{10}$$

The conditions given in (10) were commonly used in prior work (Rao and Wu, 1989; Shao, 1997). Note that these conditions are stronger than our assumptions (8)-(9), which allow high dimensional settings where $\underline{\alpha}_n^c$ varies with n. To see it, suppose that there exists $\alpha^c \in \mathcal{A}_n^c$ such that $\underline{\alpha}_n^c \not\subseteq \alpha^c$. Then some columns of X_n are linearly dependent, which contradicts the second condition in (10). This implies (8). In addition, the first three conditions in (10) imply that

$$\liminf_{n\to\infty} \min_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \Delta_n(\alpha) > 0 \quad \text{if } 0 < \text{card}(\mathcal{A}_n^c) < \text{card}(\mathcal{A}_n), \tag{11}$$

which, together with $n^{-1} d_n \to 0$, result in (9).

The loss and risk defined above may be written as

$$\mathcal{L}_n(\alpha) = \Delta_n(\alpha) + \frac{\|\boldsymbol{e}_n\|_{P_n(\alpha)}^2}{n}, \quad \mathcal{R}_n(\alpha) = \Delta_n(\alpha) + \frac{\sigma^2 d_n(\alpha)}{n}. \tag{12}$$

Define

$$\mathcal{A}_n^L = \left\{ \alpha : \alpha \in \mathcal{A}_n, \, \mathcal{L}_n(\alpha) = \min_{\alpha' \in \mathcal{A}_n} \mathcal{L}_n(\alpha') \right\}.$$

We let α_n^L denote the model with the smallest dimension in $\mathcal{A}_n^L$ (and we arbitrarily pick up one from $\mathcal{A}_n^L$ if it is not unique). When $\mathcal{A}_n^L$ contains only one element, we write $\mathcal{A}_n^L = \{\alpha_n^L\}$. We similarly define $\mathcal{A}_n^R$ and α_n^R. A model selection procedure is said to be $\mathcal{L}$-consistent if $\mathbb{P}(\hat{\alpha}_n \in \mathcal{A}_n^L) \to 1$, and asymptotically loss efficient if $\mathcal{L}_n(\hat{\alpha}_n)/\mathcal{L}_n(\alpha_n^L) \to_p 1$. Note that the former concept implies the later one. Similarly, the $\mathcal{R}$-consistency and asymptotic risk efficiency are respectively defined by $\mathbb{P}(\hat{\alpha}_n \in \mathcal{A}_n^R) \to 1$, and $\mathcal{R}_n(\hat{\alpha}_n)/\mathcal{R}_n(\alpha_n^R) \to_p 1$. We say $\mathcal{R}_n(\cdot)$ is regular, if for any sequence $\{\alpha_n\}$ that $\mathcal{R}_n(\alpha_n)/\mathcal{R}_n(\alpha_n^R) \to 1$, we have $\max_{\alpha \in \mathcal{A}_n^R}\{|d_n(\alpha_n)/d_n(\alpha) - 1|\} \to 0$.

3.2. Selection consistency and prediction efficiency

Proposition 1 gives sufficient conditions under which the consistency in selection and asymptotic efficiency are equivalent.

Proposition 1. *Suppose that $\mathcal{A}_n^c \neq \emptyset$ for all sufficiently large n.*

(i) Under conditions (8), (9), and

$$\sum_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \{n\mathcal{R}_n(\alpha)\}^{-m_1} \to 0 \qquad \text{if } \mathcal{A}_n \neq \mathcal{A}_n^c, \tag{13}$$

for some fixed constant $m_1 \geq 1$ satisfying $E(e_1^{4m_1}) < \infty$ (recall that e_1 is the noise term), we have

$$\mathbb{P}\left\{\mathcal{A}_n^R = \{\alpha_n^R\},\, \mathcal{A}_n^L = \{\alpha_n^L\},\, \alpha_n^L = \underline{\alpha}_n^c = \alpha_n^R\right\} \to 1. \tag{14}$$

(ii) If we further assume that $d_n(\underline{\alpha}_n^c) < \overline{d}$ for some fixed constant $\overline{d} > 0$, and

$$\liminf_{n\to\infty} \mathbb{P}\left\{\min_{\alpha \in \mathcal{A}_n^c \setminus \{\underline{\alpha}_n^c\}} \|\boldsymbol{e}_n\|^2_{P_n(\alpha) - P_n(\underline{\alpha}_n^c)} > \delta\right\} > 0 \tag{15}$$

for some fixed constant $\delta > 0$, then $\mathcal{L}$-consistency and $\mathcal{R}$-consistency, and asymptotic loss and risk efficiency are all equivalent.

We note that (13) is a regularity condition that has been commonly used to derive asymptotic results, for example in (Li, 1987, A.3) and (Shao, 1997, eq.(2.6)). Consider, for example, the typical case where $\mathcal{R}_n(\alpha) > n^{-\zeta}$ for all $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$ and a fixed $\zeta < 1$. Suppose that $\text{card}(\mathcal{A}_n \setminus \mathcal{A}_n^c)$ increases as a polynomial in n, and that the moment generating function of e_i exists, then condition (13) is met. An assumption stronger than (13) was made in (Shibata, 1981, Assumption 2). Condition (15) holds if, for example, $\boldsymbol{e}_n$ is Gaussian and $\mathcal{A}_n^c$ is a nested class such as $\mathcal{A}_n^c = \{\{1\}, \{1, 2\}, \ldots\}$.

3.3. Asymptotic performance of Generalized information criterion

It has been shown that the asymptotic performance of GIC_{λ_n} procedure (and thus many others) largely depends on the choice of λ_n (Nishii et al., 1984; Li, 1987; Rao and Wu, 1989; Shao, 1997). As a summary and extension of existing results, we provide the following proposition for GIC_{λ_n} in two representing cases: $\lambda_n = 2$ and $\lambda_n \to \infty$. In the sequel, we shall refer to the two cases as GIC_2 and GIC_{λ_n}, respectively. The proof of asymptotic loss efficiency in case (i) (assuming $\underline{\alpha}_n^c$ does not depend on n) and case (ii) were studied in Theorem 1 (i)(iii) and Theorem 2 (ii) in (Shao, 1997). Proposition 2 summarizes both loss and risk efficiency in two cases under weaker conditions.

Proposition 2. *Assume that $\hat{\sigma}_n^2 \to_p \sigma^2$ and conditions (8), (9), (13) hold.*

(i) In the case that $\mathcal{A}_n^c \neq \emptyset$ for all sufficiently large n (parametric), if

we further assume that

$$\limsup_{n\to\infty} \max_{\alpha\in\mathcal{A}_n\backslash\mathcal{A}_n^c} \frac{\mathcal{R}_n(\underline{\alpha}_n^c)}{\mathcal{R}_n(\alpha)}\lambda_n < 1, \tag{16}$$

$$\limsup_{n\to\infty} \sum_{\alpha\in\mathcal{A}_n^c\backslash\{\underline{\alpha}_n^c\}} \{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_2} < \infty \tag{17}$$

for some fixed constant $m_2 \geq 1$ *satisfying* $E(e_1^{4m_2}) < \infty$*, then* GIC_{λ_n} *with* $\lambda_n \to \infty$ *is* $\mathcal{L}$*-consistent and* $\mathcal{R}$*-consistent.*

(ii) In the case that card$(\mathcal{A}_n^c) \leq 1$ *for all sufficiently large* n *(parametric with a unique correct candidate model or nonparametric), then* GIC_2 *is asymptotically loss and risk efficient.*

Remark 1 (Interpretation of each condition). *Condition (16) requires* λ_n *to be not too large so that the selection procedure does not underfit. If for case (i) we assume the stronger condition (10) (thus (11)), then (16) can be implied from* $\limsup_{n\to\infty} n^{-1}\lambda_n d_n(\underline{\alpha}_n^c) = o(1)$*. If we further assume a fixed* $d_n(\underline{\alpha}_n^c)$ *that does not depend on* n*, then it suffices to require* $\lambda_n = o(n)$*, as is often assumed in the classical model selection papers.*

Assumption (17) is a regularity condition that implicitly controls the number of low-dimensional true models, so that the minimal one $\underline{\alpha}_n^c$ *can be distinguished between the remaining ones in* $\mathcal{A}_n^c$*. For nested models in the form of* $\mathcal{A}_n^c = \{\alpha_1, \alpha_2, \ldots\}$*,* $\alpha_1 \subsetneq \alpha_2 \subsetneq \cdots$*, (17) is trivially satisfied given any* $m_2 > 1$*.*

It is assumed that $\hat{\sigma}_n^2 \to_p \sigma^2$ *in both cases. A popular choice is to let* $\hat{\sigma}_n^2 = S(\bar{\alpha}_n)/(n-d_n)$*. It is consistent for* σ^2 *if* $\mathcal{A}_n^c \neq \emptyset$*, but not necessarily so if* $\mathcal{A}_n^c = \emptyset$*. For* $\mathcal{A}_n^c = \emptyset$*, it has been proved in (Shao, 1997, Thm.1A) that the additional assumptions* $\Delta_n(\bar{\alpha}_n) \to 0, d_n/n \not\to 1$ *guarantee that* $\hat{\sigma}_n^2 \to_p \sigma^2$*.*

We note that the results of both cases allow $\underline{\alpha}_n^c$ *to vary with* n*. In that sense, what is essential to distinguish the parametric and nonparametric* *situations is not the existence of a correct model of fixed dimension, but the rate of convergence of risks in $\mathcal{A}_n^c$ and $\mathcal{A}_n \setminus \mathcal{A}_n^c$ (condition (9)).*

Proposition 2 states the asymptotic loss/risk efficiency of GIC_2 (resp. GIC_{λ_n}) in the nonparametric (resp. parametric) case. In addition, if $\mathcal{A}_n^c$ contains more than one correct model with fixed dimensions, then GIC_2 is typically not asymptotically loss efficient (Shao, 1997, Thm.1(iii)). On the other hand, GIC_{λ_n} is typically not asymptotically efficient in nonparametric situations (Shao, 1997, p.230). These will also be demonstrated by our numerical experiments.

3.4. Can we select the best model consistently in the nonparametric situations?

Related to the questions proposed in Section 1, we have discussed that selection consistency (in terms of both loss and risk) is achievable in parametric situations. Proposition 1 also shows that asymptotic efficiency is often equivalent to consistency in a parametric model class.

For nonparametric situations, is it possible to consistently identify the best model (in terms of the smallest loss/risk)? This subsection gives a negative answer under mild conditions (which are easily met under familiar model approximation errors). In other words, one could approach but not exactly achieve the optimal risk, meaning that asymptotic efficiency is a more suitable concept than selection consistency for nonparametric model classes. This understanding is important because it immediately rules out the possibility of adaptation over parametric and nonparametric situations in terms of selection of the best model, but still leaves the door open for pursuing adaptive optimal estimation of the regression function. Indeed, in the next section, we propose an adaptive method that achieves asymptotic loss/risk efficiency in both parametric and nonparametric situations.

In the following theorem and its corollary, we show under reasonable assumptions that selection consistency in nonparametric models is often unrealistic. In the following theoretical analysis, we shall focus on risk consistency due to the discussion in Subsection 3.2. Most of the state-of-the-art variable selection criteria as introduced before are based on the least squares fitting error $S_n(\alpha)$ that is defined in (1). Let $\boldsymbol{S}_{\mathcal{A}_n} \triangleq [S_n(\alpha)]_{\alpha \in \mathcal{A}_n}$ be the vector of fitting errors of each model in $\mathcal{A}_n$. Let the map $\boldsymbol{y}_n \mapsto \psi_n(\boldsymbol{y}_n, \mathcal{A}_n) \in \mathcal{A}_n$ indicate a model selection procedure. Our results are concerned with a class of selection rules satisfying the following two properties.

(P1) *Least squares sufficiency*: The selection criterion $\psi_n(\boldsymbol{y}_n, \mathcal{A}_n)$ can be written as $\phi_n(\boldsymbol{S}_{\mathcal{A}_n})$ for some mapping $\boldsymbol{S}_{\mathcal{A}_n} \mapsto \phi_n(\boldsymbol{S}_{\mathcal{A}_n}) \in \mathcal{A}_n$. In other words, the selection is only using the information from least squares error $\boldsymbol{S}_{\mathcal{A}_n}$.

(P2) *Reduction compatibility*: The selection result $\hat{\alpha}$ remains unchanged if any subset excluding $\hat{\alpha}$ is removed from $\mathcal{A}_n$ (and correspondingly $\boldsymbol{S}_{\mathcal{A}_n}$ is reduced). In other words, if $\mathcal{A}'_n \subset \mathcal{A}_n$ and $\phi_n(\boldsymbol{S}_{\mathcal{A}_n}) \in \mathcal{A}'_n$, then $\phi_n(\boldsymbol{S}_{\mathcal{A}_n}) = \phi_n(\boldsymbol{S}_{\mathcal{A}'_n})$.

A selection criterion is said to be normal, if it satisfies properties (P1) and (P2). Clearly, all the penalized selection criteria as introduced before are normal selection criteria.

Theorem 1. *Assume that the noises $e_1, \ldots, e_n$ are i.i.d. Gaussian with zero mean and variance σ^2. Suppose that the model class is nested in the form of $\mathcal{A}_n = \{\alpha_1, \ldots, \alpha_{d_n}\}$, $\alpha_j = \{1, \ldots, j\}$, and $\mathcal{A}_n^c = \emptyset$. Suppose that for all sufficiently large n,*

(i) $\Delta_n(\bar{\alpha}_n) \to 0$, $d_n/n \to 0$;

(ii) $\mathcal{R}_n(\alpha_n^R) > n^{-\zeta}$ for some fixed constants $\zeta \in (0,1)$;

(iii) *card*$(\mathcal{A}_n^R) < C$ *for some fixed constant* C*;*

(iv) $\Delta_n(\alpha)$ *depends only on* α*;*

(v) $\Delta_n(\alpha_n^R) < b$ *for some constant* b

(vi) There exists n' $(n' > n)$ *that is a function of* n *and satisfies*

$$n^{-1/2}(n - n') \to 0 \tag{18}$$

$$n^{-1/2}\{d(\alpha_n^R) - d(\alpha_{n'}^R)\} \to 0, \quad n\{\Delta_n(\alpha_n^R) - \Delta_n(\alpha_{n'}^R)\} \to \infty \tag{19}$$

$$\alpha_n^R \subsetneq \alpha_{n'}^R, \quad d(\alpha_{n'}^R) \leq d_n \tag{20}$$

for any $\alpha_n^R \in \mathcal{A}_n^R$ *and* $\alpha_{n'}^R \in \mathcal{A}_{n'}^R$.

Then asymptotic efficiency can be achieved (e.g. using GIC$_2$*), while for any normal selection criterion* ψ

$$\limsup_{n\to\infty} \mathbb{P}_n\{\psi_n(\boldsymbol{y}_n, \mathcal{A}_n) \notin \mathcal{A}_n^R\} > 0, \tag{21}$$

where $\mathbb{P}_n$ *denotes the probability under the distribution of* $\boldsymbol{y}_n$.

In particular, all the above conditions are met if the model approximation error is in the familiar form of $\Delta_n(\alpha) = c\, d(\alpha)^{-\gamma}$ *with* $c > 0, 0 < \gamma < 1$.

A more general result is summarized in Proposition 4 in the supplementary file. We note that in the limit of (21), y_n and $y_{n'}$ $(n \neq n')$ are not required to be independent, so that the result applies to either independent realizations or a single realization as n varies.

3.5. Adaptive optimality of Bridge criterion

Recall that our goal of adaptive optimal variable selection is to achieve $\mathcal{L}$-consistency and $\mathcal{R}$-consistency in parametric settings and asymptotic loss/risk efficiency in nonparametric settings (thus asymptotic efficiency in general). Theorem 2 below establishes the asymptotic optimality of our proposed Bridge criterion. Its proof is in line with the intuitions explained in Section 2.

THEOREM 2. *Assume that* $\hat{\sigma}_n^2 \to_p \sigma^2$*, conditions (8), (9), (13) hold, and either of the following set of conditions hold for all sufficiently large* n*.*

- *Case 1:* $\mathcal{A}_n^c \neq \emptyset$*,* λ_n *satisfies (16), and*
$$\frac{\lambda_n}{d_n(\underline{\alpha}_n^c)} \to \infty. \tag{22}$$
Additionally, there exists a fixed constant $m_3 \geq 1$ *such that* $E(e_1^{4m_3}) < \infty$*, and*
$$\lim_{\ell \to \infty} \limsup_{n \to \infty} \sum_{\alpha \in \mathcal{A}_n^c,\, d_n(\alpha) - d_n(\underline{\alpha}_n^c) > \ell} \{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_3} = 0, \tag{23}$$
$$\limsup_{n \to \infty} \sum_{\alpha \in \mathcal{A}_n^c,\, d_n(\alpha) - d_n(\underline{\alpha}_n^c) < k} \{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_3} < \infty \tag{24}$$
for any fixed positive integer k*.*

- *Case 2:* $\mathcal{A}_n^c = \emptyset$*,* λ_n *satisfies*
$$\lambda_n \leq \frac{q\, d_n(\alpha_n^R)}{\log d_n(\alpha_n^R)} \tag{25}$$
for any fixed constant $0 < q < 2(d_0 - 1)/d_0$*, where* $d_0 > 1$ *is a constant such that* $d_n(\alpha_n^R) \geq d_0$*, and* $\mathcal{R}_n(\cdot)$ *is regular.*

Then Bridge criterion is $\mathcal{L}$*-consistent and* $\mathcal{R}$*-consistent in the first case (*$\mathcal{A}_n^c \neq \emptyset$*), and asymptotically loss and risk efficient in the second case (*$\mathcal{A}_n^c = \emptyset$*).*

3.6. Assessment of parametricness

The next Proposition 3 shows that our proposed PI_n converges in probability to one if the model class exhibits parametricness, and to zero otherwise, under some assumptions. Experimental studies in Section B show that $\textsc{pi}_n$ does provide the intended indications. For large n, it is either close to one or not, depending on whether it is truly parametric or not. In the experimental study, we also report another definition of PI given by Liu and Yang (2011), denoted by $\textsc{pi}_n^{(2)}$. The intuition is that dropping few variables produces a significantly larger increase of fitting error in a parametric model class than in a nonparametric one. $\textsc{pi}_n^{(2)}$ was shown to converges in probability to ∞ and 1 in parametric and nonparametric scenarios, respectively.

Proposition 3. *Under the same conditions of Theorem 2, if $\mathcal{A}_n^c \neq \emptyset$ (parametric), then $\textsc{pi}_n \rightarrow_p 1$; if $\mathcal{A}_n^c = \emptyset$ (nonparametric), and we further assume that the function $\mathcal{R}_n^* : \alpha \mapsto \mathcal{R}_n(\alpha) + (\lambda_n - 2)\sigma^2 d_n(\alpha)/n$ (resp. $\mathcal{R}_n : \alpha \mapsto \mathcal{R}_n(\alpha)$) is regular and has a unique minimum α_n^* (resp. α_n^R) such that*

$$\max_{\alpha \in \mathcal{A}_n} \frac{\lambda_n d_n(\alpha)}{n\mathcal{R}_n(\alpha)} < c_1, \quad \frac{d_n(\alpha_n^*)}{d_n(\alpha_n^R)} < c_2 \tag{26}$$

for all sufficiently large n for some fixed constants $c_1 > 0$ and $0 < c_2 < 1$, then $\textsc{pi}_n \rightarrow_p 0$.

3.7. Simulated validation

PI could provide a quick indication of how much parametricness the model class exhibits. It is possible to construct hypothesis test of the model class being parametric against its alternative by bootstrapping. We introduce another way of assessing parametricness which may be more intuitive. If computational cost is not an issue, data analyst may choose to generate simulated data from the selected (and estimated) model, redo the same model selection procedure as if the simulated data is the original data, and check whether the same model can be consistently selected. This idea, referred to as simulated validation, is motivated by the idea of guided simulation or cross-examination used to compare models Li et al. (2000).

In particular, suppose that BC selects model $\hat{\alpha}_{\mathrm{BC}}$, with estimated coefficients $\tilde{\boldsymbol{\beta}}_{\hat{\alpha}_{\mathrm{BC}}}$ and regression function $\tilde{\boldsymbol{f}}_n = X\tilde{\boldsymbol{\beta}}_{\hat{\alpha}_{\mathrm{BC}}}$ for brevity. If $\hat{\alpha}_{\mathrm{BC}}$ is equal to $\bar{\alpha}_n = \{1, \ldots, d_n\}$, we expand $\mathcal{A}_n$ by introducing another column of X_n such that $X_n^{\mathrm{T}} X_n$ remains invertible. We generate data $\tilde{\boldsymbol{y}}_n$ by

$$\tilde{\boldsymbol{y}}_n = \tilde{\boldsymbol{f}}_n + \tilde{\boldsymbol{\varepsilon}}_n$$

where $\tilde{\boldsymbol{\varepsilon}}_n$ is a vector of independent Gaussian noises with zero mean and $\hat{\sigma}^2$ variance that is estimated from model $\hat{\alpha}_{\mathrm{BC}}$. BC is applied to the data and select a model $\hat{\alpha}_{\mathrm{BC},s}$. Intuitively, if the model class is parametric, the probability that $\hat{\alpha}_{\mathrm{BC},s}$ coincides with $\hat{\alpha}_{\mathrm{BC}}$ is close to one. The expected result from simulated validation is formalized by the following theoretical result.

Theorem 3. *Assume that the conditions of Theorem 2 hold (for the original observations). Additionally, we assume that for any model $\alpha \in \mathcal{A}_n$ that is not the largest model $\bar{\alpha}_n$, there exists a larger model $\alpha' \in \mathcal{A}_n$ such that $\alpha \subset \alpha'$ and card(α') = card$(\alpha) + 1$. Then* $\lim_{n\to\infty} \mathbb{P}(\hat{\alpha}_{\mathrm{BC}} = \hat{\alpha}_{\mathrm{BC},s}) = 1$ *if the model class is parametric, or* $\limsup_{n\to\infty} \mathbb{P}(\hat{\alpha}_{\mathrm{BC}} = \hat{\alpha}_{\mathrm{BC},s}) \le c$ *for some constant* $c \in (0, 1)$ *if the model class is nonparametric.*

The additional assumption made in Theorem 3 is to ensure that a nonparametric model does not exhibit strong parametricness only due to a lack of competing models (whose dimensions are comparable). It is a mild assumption which holds, for instance, when $\mathcal{A}_n^c$ is a union of nested models in the form of $\{i_1\}, \{i_1, i_2\}, \ldots$.

3.8. Discussion on the λ_n in BC

In Theorem 2, we have shown that for a wide variety of λ_n's and parametric/nonparametric situations, BC enjoys the universal optimality that

GIC cannot. It is worth noting that though BC and GIC_{λ_n} have λ_n in their expressions, they lead to fundamentally different asymptotic consequences. As we discussed in Subsection 3.3, asymptotic efficiency of GIC_{λ_n} is only possible for either $\lambda_n \to \infty$ or $\lambda = 2$, depending whether the model class is parametric or not. For bounded λ_n that is not equal to 2, GIC_{λ_n} is typically sub-optimal. On the other hand, with a wide range of λ_n, BC can be optimal in both parametric or nonparametric situations. It implies that for a fairly chosen λ_n data analysts do not need to worry about whether the model class is well specified or not for optimal estimation.

A natural concern is how to select λ_n in practice, since the universal optimality of BC only holds under some conditions. From Theorem 2, these conditions depend on how parametric or nonparametric the underlying data generating model is. One way of thinking about the issue is to simply choose a deterministic sequence λ_n in advance (when the data have not been observed). Then nature flips a coin, generating a set of data in a way either parametric or nonparametric. And there will be a range of parametric or nonparametric data generating models that admit the optimal property of BC. In our simulation, we find that $\lambda_n = n^{1/3}$ is a reasonable choice in various situations. Another way is to select a data-driven λ_n using cross-validation, which will be discussed in Subsection B.4.

4. Use of Bridge criterion in high dimension ($n < d_n$)

4.1. Penalized regression

In the previous sections, we have considered variable selection in high dimensional situations where models in $\mathcal{A}_n$ and their dimensions can vary with n, but under $n \geq d_n$. Finding the sparse solution $\boldsymbol{\beta}$ of the high dimensional regression $f(\boldsymbol{x}) = \sum_{j=1}^{d_n} \beta_j x_j$ where $n < d_n$ has also received enormous attention in the past decades. The state-of-art approach is to solve a penalized regression. Commonly used penalty functions include least absolute shrinkage and selection operator (LASSO) and its extentions (Tibshirani, 1996; Zhao and Yu, 2006; Zou, 2006), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010). Given that the model is parametric and the true $\boldsymbol{\beta}$ is a sparse vector (with s nonzero entries, $s < n$), suitable conditions for consistent variable selection/estimation or prediction error bounds have been established for the above methods. More details can be found in (Fan and Li, 2001; Zhang, 2010; Hastie et al., 2015).

Nevertheless, it is not clear whether the regularization parameters usually chosen by cross validation are optimal. The theoretical performance of LASSO etc. in nonparametric situations is not fully understood. Moreover, even if the model is parametric, it can exhibit nonparametricness when the sparsity s is large relative to n (as we shall see from the experiments in Subsection B.3). Indeed, instability of these penalized regression methods is well-known and the high uncertainty damages reproducibility of the statistical findings (Meinshausen and Bühlmann, 2010; Yu, 2013; Nan and Yang, 2014; Lim and Yu, 2016). Here we propose a more reliable selection approach that takes advantage of BC together with other considerations.

4.2. New section method based on BC with variable ordering (BC-VO)

The idea is to utilize penalized regression methods to generate promising models, obtain a stable marginal weighting of the variables, and form a set of nested subsets. After we turn subset selection problem to order selection of nested models, BC can be applied and is expected to work well adaptively. We call this method BC with variable ordering in weight,

Algorithm 1 BC-VO method

```
(a). Randomly split the data into two disjoint parts $\mathcal{D}_1, \mathcal{D}_2$;
(b). Apply LASSO, SCAD, and MCP to a data set $\mathcal{D}_{1a} \subseteq \mathcal{D}_1$, and obtain the union of the three solution paths (denoted by $\hat{\mathcal{A}}_n$);
(c). Calculate the weight of each model in $\hat{\mathcal{A}}_n$ (denoted by $\boldsymbol{w} = [w_\alpha]_{\alpha \in \hat{\mathcal{A}}_n}$) from a data set $\mathcal{D}_{1b} \subseteq \mathcal{D}_1$, using an appropriate weighting scheme (described below);
(d). Calculate the marginal importance from $\boldsymbol{w}$ for each variable $k \in \{1, \ldots, d_n\}$, defined by $u_k = \sum_{\alpha \in \hat{\mathcal{A}}_n} w_\alpha 1_{k \in \alpha}$ for $k \in \cup_{\alpha \in \hat{\mathcal{A}}_n} \alpha$, and $u_k = 0$ otherwise, where $1_{k \in \alpha}$ is the 0-1 indicator;
(e). Collect the variables with nonzero $u_k$'s, arrange them in descending order of $u_k$, denoted by $i_1, i_2, \ldots$, and then form a nested model class $\mathcal{A}_n = \{\{i_1\}, \{i_1, i_2\}, \ldots\}$ of size no larger than $n$;
(f). Apply BC to the second subset of data $\mathcal{D}_2$ and select the optimal one from $\mathcal{A}_n$. Apply least squares to the selected variables for future prediction.
```

or BC-VO. Details are summarized in Algorithm 1.

The above proposed method can be applied to regression variable selection when there is no prescribed candidate models, or when $d_n > n$. Next, we explain more details of each step.

The prescreening step in (b) is important because for $d_n > n$, even the simple nested candidate set $\{\alpha_k : k = 1, \ldots, d_n\}$ with $\alpha_k = \{1, \ldots, k\}$ violates the invertibility of $X_n(\alpha)^{\mathrm{T}} X_n(\alpha)$. The $\boldsymbol{w}$ in Step (c) is introduced to measure the accuracies of the selected candidate models. To obtain $\boldsymbol{w}$, we use the adaptive regression by mixing (ARM) weighting scheme (Yang, 2001; Nan and Yang, 2014). A short review of ARM is included in Section K.1 of the supplementary file. In step (b) and (c), we will assume $\mathcal{D}_{1a}$ and $\mathcal{D}_{1b}$ to be disjoint for technical convenience. But we found from various experimental studies that $\mathcal{D}_{1a} = \mathcal{D}_{1b} = \mathcal{D}_1$ works very well in practice, and this is the default option developed in the current package 'bc'. Step (d) uses a weighting scheme to produces the ordering of variables. It was originally proposed by Ye et al. (2018) to measure variable importance in high-dimensional regression. Step (e) forms a candidate class for the application of BC.

Alternative solutions to penalized regression in Step (b) are greedy algorithms, which make locally optimal choices in each iteration and build up a nested $\hat{\mathcal{A}}_n$. Examples are the orthogonal matching pursuit (OMP) (Chen et al., 1989; Pati et al., 1993), regularized orthogonal matching pursuit (Needell and Vershynin, 2009), compressive sampling matching pursuit (Needell and Tropp, 2009), and subspace pursuit (Dai and Milenkovic, 2009). For example, the OMP algorithm is capable of identifying $\underline{\alpha}_n^c$ at iteration $d_n(\underline{\alpha}_n^c)$ with high probability, given that X is a Gaussian or Bernoulli random matrix with $n \sim s^2 \log(d_n/s)$ where s denotes the sparsity level (Davenport and Wakin, 2010). The convergence rate of OMP, a consistent model selection procedure along its solution path, and its oracle property were also studied by Ing and Lai (2011); Ding et al. (2013).

A simpler version of the above selection procedure is to apply only Steps (a)(b)(f) with $\mathcal{A}_n = \hat{\mathcal{A}}_n$. Splitting the complete data into two disjoint part (one for selecting the $\mathcal{A}_n$ and the other for α) is necessary, as we shall explain in the next remark. The idea of forming a candidate set using the solution paths of LASSO, SCAD, and MCP was also used to design a variable selection diagnostics measures for high dimensional regression (Nan and Yang, 2014).

From synthetic data experiments, we have observed that the simpler procedure without (c)-(e) works reasonably well in practice, but there can be uncertainty arising from the choice of splitting ratio. Another issue is that the final result depends the solution path returned from LASSO etc., which can be sensitive to data size (Nan and Yang, 2014) (especially when the size is small). Since variable selection is also of interest (instead of pure prediction), we propose the additional Steps (c)-(e) to alleviate the above mentioned issues. These steps first assess the accuracies of the selected candidate models, and then stabilize variable selection by assigning importance weights to variables and re-formulating the candidate set.

Our experiments (not reported due to space limitation) show that the performance of the procedure (a)-(f) is not sensitive to the splitting ratio in (a). In the last step of ARM, e^{-C_α} is introduced as a prior weight to accommodate high-dimensional settings (as found in our experiments). We refer to (Nan and Yang, 2014) for an information theoretical interpretation of C_α.

We note that the data splitting used in Steps (b) and (f) are necessary for consistency, because otherwise the procedures of candidate prescreening and model selection are not independent, invalidating the independence assumption of $e_1, \ldots, e_n$. Moreover, the arguments of asymptotic efficiency in the nonparametric scenario can be directly applied to the second subset of data. But it is not clear how to achieve asymptotic efficiency for the complete design matrix.

In the high-dimensional setting, we would like to study the prediction consistency and efficiency in both parametric and nonparametric regression models, just like their low-dimensional counterparts. However, we have not found a good way of formalizing the prediction efficiency for high-dimensional models. We leave it as an interesting future work. In Section K of the supplementary file, we only provide theoretical study for parametric high-dimensional regression models, in the sense that the regression function is a linear function of few significant variables.

Suppose that the data generating model is

$$\boldsymbol{y}_n = X_n \boldsymbol{\beta}_* + \boldsymbol{e}_n \tag{27}$$

where $\boldsymbol{\beta}_*$ has s nonzero entries, s being a fixed positive integer. Accordingly, there exists a sparse subset $\underline{\alpha}^c \in \{1, \ldots, d_n\}$ of cardinality s that represents the smallest correct model. The model $\underline{\alpha}^c$ is uniquely identifiable under some assumptions to be made in our theorem. We note that the assumption of s being fixed was not needed in our low-dimensional settings in the previous sections. We shall prove that the BC-VO estimate from step (f) exhibits "oracle property" (Fan and Li, 2001) which also holds for SCAD (Fan and Li, 2001), MCP (Zhang, 2010), Adaptive LASSO (Zou, 2006) under suitable assumptions. This property means that as the sample size and model dimension go to infinity, all and only the true variables will be identified with probability going to one, the estimated parameters converge in probability to the true parameters, and the usual asymptotic normality holds as if all the irrelevant variables have already been excluded. This property also holds for SCAD, MCP, Adaptive LASSO (Fan and Li, 2001; Zhang, 2010; Zou, 2006).

5. Conclusion

In this paper, we introduced a new variable selection criterion that achieves asymptotic loss and risk efficiency in both parametric and nonparametric situations. The proposed method is theoretically intriguing and practically useful, as it bridges the gap between two typical classes of model selection methods. Its intrinsic adaptivity to data is mainly due to the harmonic penalty $\lambda_n(1 + 2^{-1} + \cdots)$. In practice when no prior knowledge about the model specification or data generating process is available, the proposed method is more flexible and reliable than the two typical classes of criteria in selecting the most appropriate model. We also proposed a procedure for regression variable selection where the candidate set is not prescribed, or where the number of candidate variables is large relative to the sample size.

We conclude this paper by providing several ideas for further research. Firstly, since the proposed criterion admits a wide range of choices of λ_n, it may lead to a better performance by choosing data-driven λ_n in a more principled manner. Secondly, the theoretical results in this paper are for fixed design regressions, and their extension to random design analysis would be interesting. Thirdly, adaptive optimal selection of variables in very high dimensional settings has yet to be systematically studied. For instance, it is theoretically unclear whether the proposed procedure in Section 4 can achieve universal optimality as its low dimensional counterpart. Last but not least, it would be appealing to have Bayesian counterpart of the new criterion.

6. Supplementary Material

Detailed proofs and explanations of each assumption are elaborated in the supplementary file.

References

Adell, J. A. and P. Jodrá (2006). Exact Kolmogorov and total variation distances between some familiar discrete distributions. *J. Inequal. Appl. 2006*(1), 64307.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math. 21*(1), 243–247.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math. 22*(1), 203–217.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Bally, V., L. Caramellino, et al. (2016). Asymptotic development for the clt in total variation distance. *Bernoulli 22*(4), 2442–2485.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory 44*(6), 2743–2760.

Barron, A., Y. Yang, and B. Yu (1994). Asymptotically optimal function estimation by minimum complexity criteria. In *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, pp. 38. IEEE.

Breheny, P. and S. Lee (2011). Regularization paths for SCAD and MCP penalized regression models. *available at http://cran.r-project.org/web/packages/ncvreg/ncvreg.pdf*.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika 76*(3), 503–514.

Candes, E. J. and T. Tao (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory 51*(12), 4203–4215.

Casella, G., F. J. Girón, M. L. Martínez, and E. Moreno (2009). Consistency of bayesian procedures for variable selection. *Ann. Stat.*, 1207–1228.

Chen, S., S. A. Billings, and W. Luo (1989). Orthogonal least squares methods and their application to non-linear system identification. *Int. J. Control 50*(5), 1873–1896.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

Dai, W. and O. Milenkovic (2009). Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inf. Theory 55*(5), 2230–2249.

Davenport, M. A. and M. B. Wakin (2010). Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Trans. Inf. Theory 56*(9), 4395–4401.

Ding, J., L. Chen, and Y. Gu (2013). Perturbation analysis of orthogonal matching pursuit. *IEEE Trans. Signal Process. 61*(2), 398–410.

Ding, J., V. Tarokh, and Y. Yang (2018). Bridging AIC and BIC: a new criterion for autoregression. *IEEE Trans. Inf. Theory 64*(6), 4024–4043.

Durrett, R. (2010). *Probability: theory and examples.* Cambridge university press.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. *Ann. Stat. 32*(2), 407–499.

Erven, T. v., P. Grünwald, and S. De Rooij (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the aic–bic dilemma. *J. R. Stat. Soc. Ser. B 74*(3), 361–417.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Fuchs, J.-J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. Inf. Theory 51*(10), 3601–3608.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc. 70*(350), 320–328.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B 41*(2), 190–195.

Hansen, M. H. and B. Yu (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc. 96*(454), 746–774.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the LASSO and generalizations.* CRC Press.

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Ann. Stat. 38*(4), 2282.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica*, 1603–1618.

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Stat. 35*(3), 1238–1277.

Ing, C.-K. and T. L. Lai (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica*, 1473–1513.

Li, K.-C. (1987). Asymptotic optimality for Cp, Cl, cross-validation and generalized cross-validation: discrete index set. *Ann. Stat.*, 958–975.

Li, K.-C., H.-H. Lue, and C.-H. Chen (2000). Interactive tree-structured regression via principal hessian directions. *J. Amer. Statist. Assoc. 95*(450), 547–560.

Lim, C. and B. Yu (2016). Estimation stability with cross-validation (escv). *J. Comput. Graph. Statist. 25*(2), 464–492.

Liu, W. and Y. Yang (2011). Parametric or nonparametric? a parametricness index for model selection. *Ann. Stat.*, 2074–2102.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics 15*(4), 661–675.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *J. Roy. Statist. Soc. Ser. B 72*(4), 417–473.

Nan, Y. and Y. Yang (2014). Variable selection diagnostics measures for high-dimensional regression. *J. Comp. Graph. Stat. 23*(3), 636–656.

Needell, D. and J. A. Tropp (2009). Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal. 26*(3), 301–321.

Needell, D. and R. Vershynin (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math. 9*(3), 317–334.

Nishii, R. et al. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Stat. 12*(2), 758–765.

Pati, Y. C., R. Rezaiifar, and P. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *27th Asilomar Conf. Signals, Systems, Computers*, pp. 40–44. IEEE.

Prohorov, Y. V. (1952). A local theorem for densities. In *Doklady Akad. Nauk SSSR (NS)*, Volume 83, pp. 797–800.

Rao, R. and Y. Wu (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika 76*(2), 369–374.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Stat.*, 1080–1100.

Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. *Ann. Stat.*, 1012–1030.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc. 88*(422), 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica 7*(2), 221–242.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*(1), 45–54.

Speed, T. and B. Yu (1993). Model selection and prediction: normal regression. *Ann. Inst. Stat. Math. 45*(1), 35–54.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Stat.*, 689–705.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. R. Stat. Soc. Ser. B*, 44–47.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. R. Stat. Soc. Ser. B*, 276–278.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B*, 267–288.

Tibshirani, R. J. et al. (2013). The LASSO problem and uniqueness. *Electron. J. Stat. 7*, 1456–1490.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (LASSO). *IEEE Trans. Inf. Theory 55*(5), 2183–2202.

Wei, C.-Z. (1992). On predictive least squares principles. *Ann. Stat.*, 1–42.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl. 5*(3), 302–305.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc. 96*(454), 574–588.

Yang, Y. (2007). Prediction/estimation with simple linear models: is it really that simple? *Econometric Theory 23*(01), 1–36.

Yang, Y. and A. R. Barron (1998). An asymptotic property of model selection criteria. *IEEE Trans. Inf. Theory 44*(1), 95–116.

Ye, C., Y. Yang, and Y. Yang (2018). Sparsity oriented importance learning for high-dimensional linear regression. *J. Amer. Statist. Assoc.*, 1–16.

Yu, B. (2013). Stability. *Bernoulli 19*(4), 1484–1500.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 894–942.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Stat.*, 299–313.

Zhang, Y. and Y. Yang (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics 187*(1), 95–112.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *J. Mach. Learn. Res. 7*(Nov), 2541–2563.

Zhou, S., X. Shen, and D. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *Ann. Stat. 26*(5), 1760–1782.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc. 101*(476), 1418–1429.

Web-based supporting materials for "Optimal variable selection in regression models" by Jie Ding, Vahid Tarokh, Yuhong Yang

A. Notation

We use → and $\to_p$ to denote respectively deterministic convergence and in probability convergence, as the sample size $n \to \infty$. We may drop the $n \to \infty$ whenever there is no ambiguity. Let $\mathrm{eig}_{\min}(\cdot)$ and $\mathrm{eig}_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a symmetric matrix. Let $\mathrm{card}(A)$, $\lfloor a \rfloor$, and $H_n = \sum_{k=1}^{n} k^{-1}$ denote respectively the cardinality of a finite set, the largest integer that is no larger than a, and the n-th harmonic number. We use $A \setminus B$ to represent the set of elements in set A but not in set B. For a positive semidefinite matrix P, let $\|\cdot\|_P$ denote the norm defined by $\|\boldsymbol{e}\|_P^2 \overset{\Delta}{=} \boldsymbol{e}^{\mathrm{T}} P \boldsymbol{e}$. It reduces to the Euclidean norm $\|\cdot\|$ when P is an identity matrix.

B. Experimental studies

In this section, we present experimental results to demonstrate the theoretical results and the advantages of Bridge criterion on various synthetic data. The main purpose of the experiments is to demonstrate that the proposed criterion achieves asymptotic efficiency in both parametric and nonparametric settings (which is unknown beforehand). We also numerically study the performance in high dimensional regression models where $d_n > n$.

In Subsections B.1 and B.2, the model classes under consideration are respectively polynomial regression and spline regression. In both cases, the candidate set is naturally chosen to be $\{\{1\}, \ldots, \{1, \ldots, d\}, \ldots\}$, with d denoting either the degree of polynomial or number of basis splines being used. The purpose is to show that as sample size increases, BC behaves closer to the better of GIC_2-type or GIC_{λ_n}-type criteria. In Subsection B.3, we extend the numerical study of BC to high dimensional regressions where $n < d_n$, by using the procedure proposed in Section 4. Finally, in Subsection 3.8, we briefly discuss the choice of λ_n in practice.

In some of the tasks, we compare GIC_2, CV_1, BC, GIC_{λ_n}, CV_d in terms of the loss, risk, and selected dimension. The λ_n in GIC_{λ_n} is chosen to be $\log n$ so that GIC_{λ_n} is almost BIC. The d in CV_d is chosen to meet (4) with λ_n being $\log n$, so that CV_d is also comparable to BIC. In applying CV_d procedure, for each candidate model, $n - d$ randomly chosen observations are used for training and an average predication error is computed using the remaining d observations; then the prediction error is further averaged over many independent replications, based on which the optimal candidate is selected.

Throughout the experiments, the λ_n in BC is chosen to be $\lambda_n = n^{1/3}$, except in Subsection B.4 an adaptive choice of λ_n is compared. Better results may be obtained by fine tuning the optimal λ_n for each data generating scheme. But we found from our experiments that $\lambda_n = n^{1/3}$ gives a reasonable baseline level of performance without the need to do any further tuning. Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution with mean μ and variance σ^2. The noises e_i's are generated from i.i.d. $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$ unless otherwise stated. The signal to noise ratio (SNR) is defined by $\|\boldsymbol{f}_n\|^2/(n\sigma^2)$. In a parametric model class $\mathcal{A}_n$, for each $\alpha \in \mathcal{A}_n$ we define $u_n^+(\alpha) = \text{card}(\alpha \backslash \underline{\alpha}_n^c)$, $u_n^-(\alpha) = \text{card}(\underline{\alpha}_n^c \backslash \alpha)$, which represent the extent of overfitting and underfitting, respectively. In the sequel, we summarize the averages of $u_n^+(\hat{\alpha}_n)$, $u_n^-(\hat{\alpha}_n)$ in parametric settings and the averages of $d_n(\hat{\alpha}_n)$ in nonparametric settings, with $\hat{\alpha}_n$ denoting the models selected by each criterion. We also show the values of two different PI's, one is the PI_n proposed in this paper, and the other is $\text{PI}_n^{(2)}$ proposed by Liu and Yang (2011), defined in Section 2. Each experiment will be based on 1000 independent replications, each of which consists of n independent data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. Each mean estimate in the tables is followed by its standard error in the parenthesis.

B.1. Polynomial regression

Suppose that the regression function is $f(x) = \sum_{j=0}^{\infty} \beta_{j+1} x^j$ ($\boldsymbol{\beta} \in \ell_2, 0 \leq x < 1$), and $y_1, \ldots, y_n$ are observed at $x = ia/n, i = 0, \ldots, n-1$ ($0 < a \leq 1$). The (i,j)th ($1 \leq i, j \leq d_n$) element of X_n is $\{(i-1)a/n\}^{j-1}$. The asymptotic property of X_n has been studied in (Shibata, 1981, Example 3.1). We generated data from model $f(x) = 1 + 5x^2$ (denoted by $\mathcal{M}_1$), and another model $f(x) = \log(1 + 46x)$ (denoted by $\mathcal{M}_2$). Clearly, under the specified (polynomial regression) model class, $\mathcal{M}_1$ is parametric while $\mathcal{M}_2$ is nonparametric. The SNRs are 9.3 in both cases. The candidate models are chosen be $\mathcal{A}_n = \{\alpha_k, k = 1, \ldots, d_n\}$ with $d_n = \lfloor n^{1/3} \rfloor$, where α_k corresponds to $f(x) = \sum_{j=0}^{k-1} \beta_{j+1} x^j$. The dimension of model α_k is k.

We summarize the performances of BC (employing the default $\lambda_n = n^{1/3}$), GIC$_2$, CV$_1$, GIC$_{\lambda_n}$, and CV$_d$ for each $n = 100, 500, 1000$ in Table 1 and Table 2. The values of $\mathcal{L}_n(\hat{\alpha}_n)$, $\mathcal{R}_n(\hat{\alpha}_n)$ are analytically calculated using (12). In Table 1, GIC$_{\lambda_n}$ and CV$_d$ (the "first class") perform better than GIC$_2$ and CV$_1$ (the "second class") for large n $(500, 1000)$, while the other way around for small n (100). That is because when n is small relative to $d_n(\underline{\alpha}_n^c)$, it falls into the practically nonparametric regime. In Table 2, the second class outperform the first class for each n. Nevertheless, in all cases the performance of BC is close to the better of the two classes of criteria.

B.2. Spline fitting

In this subsection, we use B-splines to approximate an unknown scalar function $f(x)$, given samples of x and its noisy observations $y = f(x) + \varepsilon$. Similar technique may be extended to multivariate x by using multidimensional splines or assuming additive nonlinear models (Stone, 1985; Huang et al., 2010). The specified model class is $f(x) = \sum_{j=1}^{k} \beta_j B_{\ell,j}(x)$, where $B_{\ell,j}(x)$ are B-spline functions of order ℓ with knot sequence $t_0 = \cdots = t_{\ell-1} = 0, t_k = \cdots = t_{k+\ell-1} = 1$, $t_\ell, \ldots, t_{k-1}$ being internal knots equally spaced in $[0, 1]$. The corresponding design matrix X_n of size $n \times d_n$ is thus defined by $X_{i,j} = B_{\ell,j}(x_i)$, $i = 1, \ldots, n$. It is known that X_n is of full rank given that $\sup_{x\in[0,1]} |Q_n(x) - Q(x)| = o(k^{-1})$, where $Q_n(x)$ is the empirical distribution function of $\{x_i\}_{i=1}^n$, and $Q(x)$ is some distribution with a positive continuous density (Zhou et al., 1998, Lemma 6.2). Note that by Donsker's theorem, the above condition holds with probability close to one for x_i's being generated from $Q(x)$ and $k = o(n^{0.5})$.

We use quadratic splines ($\ell = 3$) to fit synthetic data that are generated from two models, with $x_i = i/n$, $i = 1, \ldots, n$. The parametric model (denoted by $\mathcal{M}_3$) uses $f(x) = -40x^2 + 40x - 4$. The nonparametric model (denoted by $\mathcal{M}_4$) uses $f(x) = \log(1 + 100x)$ for $x < 0.7$ and $\log(71)\{1 + 100(x - 0.7)^4\}$ for $x \geq 0.7$. The SNRs are 16 in both cases. The candidate models are chosen to be $\mathcal{A}_n = \{\alpha_k, k = 1, \ldots, d_n\}$ with $d_n = \lfloor n^{1/3} \rfloor$, where α_k corresponds to the model of k spline basis. The dimension of α_k is defined to be k. The results are summarized in Table 3 and Table 4. In Table 3, GIC_{λ_n} and CV_d perform better than GIC_2 and CV_1 for each n. As n increases, BC behaves closer to the first class, and PI_n is closer to one. In Table 4, however, the second class perform better, and so does BC.

B.3. Variable selection in high dimensional models with $n < d_n$

In this numerical study, we consider the variable selection for $f(\boldsymbol{x}) = \sum_{j=1}^{d_n} \beta_j x_j$ with $d_n > n$. We adopt the procedure BC-VO proposed in Section 4 that casts the high dimension problem as a low dimensional one for BC. Since random design is more commonly used than fixed design in high dimensional problems, it is natural to evaluate the predictive performance based on the following risk. $\mathcal{R}_n(\hat{\alpha}_n) = E(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\hat{\alpha}_n) \rangle)^2$ where $\hat{\alpha}_n$ is the selected model by a criterion, and the expectation is taken over $\hat{\boldsymbol{\beta}}(\hat{\alpha}_n)$ and covariate $\tilde{\boldsymbol{x}}$ that is independent of $\hat{\boldsymbol{\beta}}(\hat{\alpha}_n)$. We note that our present theoretical analysis for asymptotic efficiency cannot be directly applied to the random design regression, but the consistency arguments (see Section 4) are still applicable.

For each synthetic data in the sequel, covariates (rows of X_n) are independently generated. We numerically compute the mean and standard error of $\mathcal{R}_n(\hat{\alpha}_n)$ in the following way. In each replication, suppose that $\hat{\boldsymbol{\beta}}(\hat{\alpha}_n)$ is the estimated coefficients by a criterion, then we compute the average of $\langle \tilde{\boldsymbol{x}}_i, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\hat{\alpha}_n) \rangle)^2$ for randomly generated $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{1000}$. We split the data with ratio $r = 0.7$ when applying step (a) of the procedure in Section 4. In some experiments not reported here, we also chose different r and the results in Table 5 did not differ much. We use the R package NCVREG (Breheny and Lee, 2011) to perform LASSO, SCAD, and MCP model selections, and they are applied to the complete dataset (for a fair comparison with other criteria). The default 10-fold cross-validation has been used to tune the regularization parameters. Let σ denote the standard deviation of the zero mean Gaussian noise added to the correct linear model. We consider the three models described below, each with $n = 150, 450$, $d_n = 4n/3$, $\sigma = 1.5, 4.5$.

Model $\mathcal{M}_5$. In the first model, $\boldsymbol{\beta} = [10, 5, 5, 2.5, 2.5, 1.25, 1.25, 0.675, 0.675, 0.3125,$ $0.3125, 0, \ldots, 0]$ (with $d_n - 11$ zeros), and the d_n covariates are i.i.d. $\mathcal{N}(0, 1)$. The SNR is around 73 for $\sigma = 1.5$ and 6.5 for $\sigma = 5$. This model was used by Nan and Yang (2014).

Model $\mathcal{M}_6$. In the second model, $\boldsymbol{\beta} = [2.5, \ldots, 2.5, 1.5, \ldots, 1.5, 0.5, \ldots, 0.5, \ldots, 0,$ $\ldots, 0]$ (where 2.5, 1.5, 0.5 are each repeated five times, followed by $d_n - 15$ zeros). The first 15 covariates and the remaining 185 covariates are independent. The pairwise covariance between x_i, x_j is $0.5^{|i-j|}$ for $1 \leq i \leq j \leq 15$, and x_i's are i.i.d. $\mathcal{N}(0, 1)$ for $16 \leq i \leq 200$. The SNR is around 52 for $\sigma = 1.5$ and 4.5 for $\sigma = 5$. This model was used by Huang et al. (2008).

Model $\mathcal{M}_7$. In the third model, $\boldsymbol{\beta} = [10.5, 0, \ldots, 0]$ (with $d_n - 1$ zeros), and the d_n covariates are i.i.d. $\mathcal{N}(0, 1)$. The SNR is around 50 for $\sigma = 1.5$ and 4.3 for $\sigma = 5$.

Model $\mathcal{M}_8$. In the fourth model, $\boldsymbol{\beta} = [10, 10/2^{1.5}, \ldots, 10/p^{1.5}]$ (with no zeros), and the d_n covariates are i.i.d. $\mathcal{N}(0, 1)$. The SNR is around 53 for $\sigma = 1.5$ and 4.8 for $\sigma = 5$. This model appears to be nonparametric but can exhibit strong parametricness as sample size increases.

We first consider $\sigma = 1.5$. Table 5 summarizes the results in terms of $\mathcal{R}_n(\hat{\alpha}_n)$, $u_n^+(\hat{\alpha}_n)$, $u_n^-(\hat{\alpha}_n)$, and the smallest risk in each setting is bolded. For the four models, though the data are truly generated from linear models, they exhibit different practical parametricness. In $\mathcal{M}_5$ and $\mathcal{M}_6$, BC-VO substantially outperforms the state-of-art penalized methods LASSO, SCAD, and MCP, especially for larger n. Note that compared with BC, both SCAD and MCP suffer less underfitting as well as overfitting, but their risks are larger. To see the reason, we decompose the risk into bias and variance terms in the following way:

$$E(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \rangle)^2 = E_{\tilde{\boldsymbol{x}}}(\langle \tilde{\boldsymbol{x}}, \boldsymbol{\beta} - E\hat{\boldsymbol{\beta}} \rangle)^2 + E_{\tilde{\boldsymbol{x}}} Var\{\tilde{\boldsymbol{x}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}\}$$

where $\hat{\boldsymbol{\beta}}$ denotes the estimated coefficient vector, and inside $E_{\tilde{\boldsymbol{x}}}$ is the expectation/variance with respect to $\hat{\boldsymbol{\beta}}$ (given $\tilde{\boldsymbol{x}}$). Our experiments (not reported due to space limitation) show that the larger risks of of SCAD and MCP are due to their larger bias (than BC) in estimating the coefficients.

For larger n, the model $\mathcal{M}_5$ or $\mathcal{M}_6$ exhibits more parametricness (since the minimal true model does not vary with n), and thus PI_n should become larger. This is consistent with the PI's in Table 5. As we discussed before, a correct model with larger SNR and smaller dimension tends to be more "parametric". This seems also true in the high dimensional setting. To further illustrate that point, in $\mathcal{M}_7$, we reduce the dimension of the data generating model while keeping the SNR close to $\mathcal{M}_6$. The value of PI_n is close to one, and SCAD, MCP, BC-VO perform similarly for both n. Though LASSO does not exhibit much overfitting, it gives a large risk due to biased estimation of the ℓ_1-penalization. In contrast to $\mathcal{M}_7$, model $\mathcal{M}_8$ is a nonparametric setting. As sample size becomes larger, the SNR is saturated and the parametricness becomes more evident. In these four experiments, the performance of BC-VO is much better than the state-of-art penalized methods in more practically nonparametric models or comparable to them in more parametric models.

We then consider $\sigma = 5$ to further study low SNR situations. We repeat all the above experiments and show the results in Table 7. From the simulation results, there is no method dominantly better than the others. BC-VO seems to underfit for very low SNR. We found that its performance can be improved by choosing smaller λ_n. For the tabulated results, we have used the same option (namely $n^{1/3}$) to avoid cherry-picking for our own method. An SNR-adaptive choice of λ_n is an interesting future work. Overall, it seems that the performance of BC-VO is stable and comparable with the best result in almost all situations.

All the above experimental results suggest that BC-VO is a promising new method for high dimensional regression.

B.4. On the choice of λ_n in BC

In Section 3.8, we have discussed that BC can be universally optimal for a wide variety of λ_n's. In practical situations where the theoretical assumptions are not easily verifiable, we here propose a data-driven selection of λ_n (referred to as BC-Dat). The procedure proceeds as follows:

1) resample a dataset of size n (with replacement) $\mathcal{D}_1$ from the original dataset $\mathcal{D}$ to which BC should be applied,

2) for a grid of λ_n (e.g. a geometric progression between $(1, n)$ with common ratio 1.5 in our simulation), run BC, and apply the inference results to $\mathcal{D}$ to obtain a fitting error ϵ_1,

3) repeat steps (1)-(2) several times (e.g. 5 times in our simulation), and select the λ_n that offers the smallest $\epsilon_1 + \cdots + \epsilon_5$.

In the above experiments, we also compare BC-Dat and BC, and summarize their risks $\mathcal{R}_n(\hat{\alpha}_n)$ for models $\mathcal{M}_1$-$\mathcal{M}_8$ (with $\sigma = 1.5$) in Table 7. The results indicate that the performances of BC-Dat and BC are comparable, and that BC is not very sensitive to the choice of λ_n. We leave a more principled data-driven selection of λ_n (with provable guarantees) as an interesting future work.

		GIC_2	CV_1	BC	GIC_{λ_n}	CV_d
$n = 100$	$\mathcal{L}_n(\hat{\alpha}_n)$	35.21 (0.91)	35.41 (0.93)	36.16 (1.01)	37.67 (1.25)	39.35 (1.34)
	$\mathcal{R}_n(\hat{\alpha}_n)$	32.27 (0.34)	32.56 (0.38)	34.17 (0.63)	38.23 (0.98)	40.47 (1.11)
	$u_n^+(\hat{\alpha}_n)$	0.14 (0.01)	0.15 (0.01)	0.12 (0.01)	0.03 (0.01)	0.02 (0.00)
	$u_n^-(\hat{\alpha}_n)$	0.01 (0.00)	0.01 (0.00)	0.02 (0.00)	0.06 (0.01)	0.08 (0.01)
	PI_n			0.89 (0.01)		
	$\text{PI}_n^{(2)}$			3.16 (0.16)		
$n = 500$	$\mathcal{L}_n(\hat{\alpha}_n)$	8.61 (0.25)	8.60 (0.24)	6.73 (0.20)	6.10 (0.17)	6.04 (0.16)
	$\mathcal{R}_n(\hat{\alpha}_n)$	6.92 (0.06)	6.92 (0.06)	6.22 (0.03)	6.02 (0.01)	6.02 (0.01)
	$u_n^+(\hat{\alpha}_n)$	0.46 (0.03)	0.46 (0.03)	0.11 (0.02)	0.01 (0.00)	0.01 (0.00)
	$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	PI_n			0.95 (0.01)		
	$\text{PI}_n^{(2)}$			5.20 (0.04)		
$n = 1000$	$\mathcal{L}_n(\hat{\alpha}_n)$	5.08 (0.15)	5.13 (0.15)	3.49 (0.10)	3.25 (0.08)	3.25 (0.08)
	$\mathcal{R}_n(\hat{\alpha}_n)$	3.69 (0.04)	3.72 (0.04)	3.07 (0.01)	3.01 (0.00)	3.01 (0.00)
	$u_n^+(\hat{\alpha}_n)$	0.69 (0.04)	0.72 (0.04)	0.07 (0.01)	0.01 (0.00)	0.01 (0.00)
	$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	PI_n			0.98 (0.01)		
	$\text{PI}_n^{(2)}$			8.75 (0.06)		

Table 1: Performance of each method for polynomial regression in the parametric model $\mathcal{M}_1$ (values corresponding to $\mathcal{L}_n(\hat{\alpha}_n)$ and $\mathcal{R}_n(\hat{\alpha}_n)$ were rescaled by 1000)

		GIC$_2$	CV$_1$	BC	GIC$_{\lambda_n}$	CV$_d$
$n = 100$	$\mathcal{L}_n(\hat{\alpha}_n)$	58.63 (0.97)	58.80 (0.99)	62.81 (1.10)	73.79 (1.21)	78.58 (1.28)
	$\mathcal{R}_n(\hat{\alpha}_n)$	59.22 (0.54)	59.38 (0.56)	63.18 (0.74)	72.96 (0.94)	77.64 (1.01)
	$d_n(\hat{\alpha}_n)$	3.47 (0.02)	3.47 (0.02)	3.36 (0.02)	3.03 (0.02)	2.92 (0.02)
	PI_n			0.70 (0.01)		
	$\mathrm{PI}_n^{(2)}$			2.53 (0.07)		
$n = 500$	$\mathcal{L}_n(\hat{\alpha}_n)$	15.25 (0.22)	15.21 (0.22)	16.91 (0.27)	20.58 (0.29)	21.11 (0.31)
	$\mathcal{R}_n(\hat{\alpha}_n)$	14.59 (0.08)	14.56 (0.08)	16.26 (0.16)	20.21 (0.24)	20.85 (0.25)
	$d_n(\hat{\alpha}_n)$	5.26 (0.03)	5.26 (0.03)	4.82 (0.04)	3.97 (0.02)	3.89 (0.02)
	PI_n			0.62 (0.02)		
	$\mathrm{PI}_n^{(2)}$			1.64 (0.02)		
$n = 1000$	$\mathcal{L}_n(\hat{\alpha}_n)$	8.66 (0.13)	8.62 (0.14)	9.56 (0.14)	11.53 (0.16)	11.65 (0.17)
	$\mathcal{R}_n(\hat{\alpha}_n)$	8.30 (0.05)	8.28 (0.05)	9.27 (0.09)	11.43 (0.13)	11.57 (0.14)
	$d_n(\hat{\alpha}_n)$	6.07 (0.04)	6.06 (0.04)	5.34 (0.04)	4.43 (0.02)	4.40 (0.02)
	PI_n			0.56(0.02)		
	$\mathrm{PI}_n^{(2)}$			1.42 (0.02)		

Table 2: Performance of each method for polynomial regression in the nonparametric model $\mathcal{M}_2$ (values corresponding to $\mathcal{L}_n(\hat{\alpha}_n)$ and $\mathcal{R}_n(\hat{\alpha}_n)$ were rescaled by 1000)

		GIC$_2$	CV$_1$	BC	GIC$_{\lambda_n}$	CV$_d$
$n = 100$	$\mathcal{L}_n(\hat{\alpha}_n)$	36.14 (0.89)	36.10 (0.88)	35.65 (0.88)	32.43 (0.85)	32.31 (0.83)
	$\mathcal{R}_n(\hat{\alpha}_n)$	31.63 (0.12)	31.63 (0.12)	31.41 (0.11)	30.36 (0.06)	30.44 (0.06)
	$u_n^+(\hat{\alpha}_n)$	0.16 (0.01)	0.16 (0.01)	0.14 (0.01)	0.03 (0.01)	0.04 (0.01)
	$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	PI_n			0.90 (0.01)		
	$\text{PI}_n^{(2)}$			96.44 (0.58)		
$n = 500$	$\mathcal{L}_n(\hat{\alpha}_n)$	9.52 (0.26)	9.48 (0.25)	7.30 (0.22)	6.40 (0.17)	6.38 (0.17)
	$\mathcal{R}_n(\hat{\alpha}_n)$	7.16 (0.07)	7.14 (0.07)	6.26 (0.03)	6.03 (0.01)	6.03 (0.01)
	$d_n(\hat{\alpha}_n)$	1.58 (0.03)	1.57 (0.03)	1.13 (0.02)	1.02 (0.00)	1.02 (0.00)
	$u_n^+(\hat{\alpha}_n)$	0.58 (0.03)	0.57 (0.03)	0.13 (0.02)	0.02 (0.00)	0.01 (0.00)
	$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	PI_n			0.94 (0.01)		
	$\text{PI}_n^{(2)}$			488.03 (2.42)		
$n = 1000$	$\mathcal{L}_n(\hat{\alpha}_n)$	4.80 (0.14)	4.76 (0.14)	3.26 (0.10)	3.02 (0.08)	3.02 (0.08)
	$\mathcal{R}_n(\hat{\alpha}_n)$	3.63 (0.04)	3.61 (0.04)	3.07 (0.01)	3.01 (0.00)	3.01 (0.00)
	$d_n(\hat{\alpha}_n)$	3.62 (0.04)	3.65 (0.04)	3.07 (0.01)	3.01 (0.00)	3.01 (0.00)
	$u_n^+(\hat{\alpha}_n)$	0.63 (0.04)	0.61 (0.04)	0.07 (0.01)	0.01 (0.00)	0.01 (0.00)
	$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	PI_n			0.97 (0.01)		
	$\text{PI}_n^{(2)}$			987.01 (3.58)		

Table 3: Performance of each method for spline fitting in the parametric model $\mathcal{M}_3$ (values corresponding to $\mathcal{L}_n(\hat{\alpha}_n)$ and $\mathcal{R}_n(\hat{\alpha}_n)$ were rescaled by 1000)

		GIC_2	CV_1	BC	GIC_{λ_n}	CV_d
$n = 100$	$\mathcal{L}_n(\hat{\alpha}_n)$	11.56 (0.10)	11.57 (0.10)	11.56 (0.10)	12.09 (0.14)	12.62 (0.15)
	$\mathcal{R}_n(\hat{\alpha}_n)$	11.61 (0.05)	11.62 (0.05)	11.61 (0.05)	12.26 (0.11)	12.77 (0.14)
	$d_n(\hat{\alpha}_n)$	1.99 (0.00)	1.99 (0.00)	1.99 (0.00)	1.94 (0.01)	1.91 (0.01)
	PI_n			0.96 (0.01)		
	$\mathrm{PI}_n^{(2)}$			6.77 (0.68)		
$n = 500$	$\mathcal{L}_n(\hat{\alpha}_n)$	2.10 (0.01)	2.10 (0.01)	2.13 (0.03)	2.61 (0.04)	2.69 (0.04)
	$\mathcal{R}_n(\hat{\alpha}_n)$	2.10 (0.01)	2.10 (0.01)	2.12 (0.01)	2.59 (0.03)	2.69 (0.03)
	$d_n(\hat{\alpha}_n)$	4.79 (0.01)	4.79 (0.01)	4.76 (0.01)	4.22 (0.02)	4.15 (0.02)
	PI_n			0.48 (0.02)		
	$\mathrm{PI}_n^{(2)}$			1.23 (0.01)		
$n = 1000$	$\mathcal{L}_n(\hat{\alpha}_n)$	1.22 (0.01)	1.22 (0.01)	1.24 (0.01)	1.52 (0.02)	1.53 (0.02)
	$\mathcal{R}_n(\hat{\alpha}_n)$	1.19 (0.01)	1.19 (0.01)	1.20 (0.01)	1.49 (0.01)	1.50 (0.01)
	$d_n(\hat{\alpha}_n)$	5.97 (0.03)	5.98 (0.03)	5.85 (0.03)	4.74 (0.02)	4.73 (0.02)
	PI_n			0.30 (0.01)		
	$\mathrm{PI}_n^{(2)}$			1.25 (0.01)		

Table 4: Performance of each method for spline fitting in the nonparametric model $\mathcal{M}_4$ (values corresponding to $\mathcal{L}_n(\hat{\alpha}_n)$ and $\mathcal{R}_n(\hat{\alpha}_n)$ were rescaled by 100)

			LASSO	SCAD	MCP	BC
$\mathcal{M}_5$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.99 (0.02)	1.35 (0.02)	0.94 (0.01)	**0.66** (0.01)
		$u_n^+(\hat{\alpha}_n)$	0.85 (0.03)	0.12 (0.01)	0.05 (0.01)	0.73 (0.04)
		$u_n^-(\hat{\alpha}_n)$	2.19 (0.02)	2.22 (0.02)	2.22 (0.02)	2.39 (0.03)
		PI_n	0.64 (0.02)			
		$\text{PI}_n^{(2)}$	1.44 (0.02)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.61 (0.01)	1.08 (0.01)	0.69 (0.01)	**0.28** (0.01)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)
		$u_n^-(\hat{\alpha}_n)$	2.03 (0.01)	2.05 (0.01)	2.03 (0.01)	1.98 (0.01)
		PI_n	0.96 (0.01)			
		$\text{PI}_n^{(2)}$	1.56 (0.01)			
$\mathcal{M}_6$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	1.21 (0.01)	1.01 (0.02)	0.80 (0.01)	**0.67** (0.01)
		$u_n^+(\hat{\alpha}_n)$	2.00 (0.06)	1.28 (0.05)	0.88 (0.04)	0.66 (0.03)
		$u_n^-(\hat{\alpha}_n)$	0.22 (0.01)	1.12 (0.03)	2.12 (0.02)	1.74 (0.03)
		PI_n	0.62 (0.02)			
		$\text{PI}_n^{(2)}$	1.15 (0.01)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	0.80 (0.01)	0.47 (0.01)	0.46 (0.01)	**0.10** (0.00)
		$u_n^+(\hat{\alpha}_n)$	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.10 (0.01)
		$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.31 (0.02)	1.57 (0.02)	0.10 (0.01)
		PI_n	0.75 (0.01)			
		$\text{PI}_n^{(2)}$	1.29 (0.01)			
$\mathcal{M}_7$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	3.14 (0.01)	**0.37** (0.00)	**0.37** (0.00)	0.39 (0.03)
		$u_n^+(\hat{\alpha}_n)$	0.17 (0.04)	0.08 (0.03)	0.08 (0.03)	0.21 (0.05)
		$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		PI_n	1.00 (0.00)			
		$\text{PI}_n^{(2)}$	44.26 (0.18)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.78 (0.05)	**0.05** (0.05)	**0.05** (0.05)	**0.05** (0.03)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		PI_n	1.00 (0.00)			
		$\text{PI}_n^{(2)}$	135.00 (0.00)			
$\mathcal{M}_8$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.92 (0.02)	2.03 (0.01)	1.64 (0.01)	**1.45** (0.01)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		$u_n^-(\hat{\alpha}_n)$	191.92 (0.05)	192.58 (0.04)	192.79 (0.04)	192.33 (0.07)
		PI_n	0.57 (0.02)			
		$\text{PI}_n^{(2)}$	1.52 (0.02)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.71 (0.01)	1.84 (0.01)	1.47 (0.01)	**0.97** (0.01)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		$u_n^-(\hat{\alpha}_n)$	593.01 (0.03)	593.04 (0.03)	593.08 (0.02)	592.79 (0.03)
		PI_n	0.87 (0.01)			
		$\text{PI}_n^{(2)}$	1.43 (0.01)			

Table 5: Performance of each method for high dimensional models $\mathcal{M}_5$–$\mathcal{M}_8$ with $\sigma = 1.5$ (values for $\mathcal{R}_n(\hat{\alpha}_n)$, $u_n^+(\hat{\alpha}_n)$, $u_n^-(\hat{\alpha}_n)$ in $\mathcal{M}_7$ were rescaled by 10)

			LASSO	SCAD	MCP	BC
$\mathcal{M}_5$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	8.04 (0.09)	**5.21** (0.08)	5.24 (0.08)	6.65 (0.10)
		$u_n^+(\hat{\alpha}_n)$	20.48 (0.27)	12.66 (0.22)	4.78 (0.13)	2.01 (0.09)
		$u_n^-(\hat{\alpha}_n)$	3.12 (0.03)	3.41 (0.03)	4.11 (0.03)	5.28 (0.03)
		PI_n	0.52 (0.02)			
		$\mathrm{PI}_n^{(2)}$	1.46 (1.26)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	3.63 (0.05)	1.99 (0.04)	**1.84** (0.04)	2.41 (0.05)
		$u_n^+(\hat{\alpha}_n)$	19.38 (0.21)	13.78 (0.22)	6.67 (0.18)	2.23 (0.12)
		$u_n^-(\hat{\alpha}_n)$	2.23 (0.03)	2.34 (0.03)	2.73 (0.03)	3.69 (0.03)
		PI_n	0.41 (0.02)			
		$\mathrm{PI}_n^{(2)}$	1.43 (0.02)			
$\mathcal{M}_6$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	**7.15** (0.09)	11.35 (0.13)	11.53 (0.14)	9.25 (0.14)
		$u_n^+(\hat{\alpha}_n)$	18.36 (0.35)	19.10 (0.25)	8.72 (0.15)	2.62 (0.08)
		$u_n^-(\hat{\alpha}_n)$	2.31 (0.03)	4.24 (0.04)	5.86 (0.04)	5.59 (0.06)
		PI_n	0.37 (0.02)			
		$\mathrm{PI}_n^{(2)}$	1.30 (0.01)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	2.75 (0.05)	3.13 (0.05)	2.91 (0.05)	**2.42** (0.05)
		$u_n^+(\hat{\alpha}_n)$	25.55 (0.48)	30.84 (0.50)	12.10 (0.26)	2.18 (0.09)
		$u_n^-(\hat{\alpha}_n)$	0.93 (0.03)	1.87 (0.03)	3.05 (0.03)	2.92 (0.04)
		PI_n	0.29 (0.01)			
		$\mathrm{PI}_n^{(2)}$	1.20 (0.01)			
$\mathcal{M}_7$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	1.74 (0.05)	**0.56** (0.04)	0.65 (0.05)	1.25 (0.08)
		$u_n^+(\hat{\alpha}_n)$	7.12 (0.25)	2.77 (0.17)	1.46 (0.10)	1.39 (0.10)
		$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		PI_n	0.82 (0.01)			
		$\mathrm{PI}_n^{(2)}$	40.92 (0.40)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	0.66 (0.04)	**0.21** (0.04)	0.23 (0.04)	0.31 (0.04)
		$u_n^+(\hat{\alpha}_n)$	6.25 (0.20)	2.87 (0.17)	1.51 (0.11)	0.54 (0.07)
		$u_n^-(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		PI_n	0.94 (0.01)			
		$\mathrm{PI}_n^{(2)}$	130.55 (0.76)			
$\mathcal{M}_8$	$n = 150$	$\mathcal{R}_n(\hat{\alpha}_n)$	6.65 (0.06)	**5.18** (0.05)	5.23 (0.06)	6.74 (0.08)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		$u_n^-(\hat{\alpha}_n)$	178.19 (0.31)	184.23 (0.25)	191.12 (0.15)	194.45 (0.13)
		PI_n	0.55 (0.02)			
		$\mathrm{PI}_n^{(2)}$	1.49 (0.89)			
	$n = 450$	$\mathcal{R}_n(\hat{\alpha}_n)$	3.44 (0.02)	2.54 (0.02)	**2.48** (0.02)	3.18 (0.03)
		$u_n^+(\hat{\alpha}_n)$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
		$u_n^-(\hat{\alpha}_n)$	574.65 (0.23)	578.32 (0.25)	586.17 (0.22)	592.72 (0.15)
		PI_n	0.41 (0.14)			
		$\mathrm{PI}_n^{(2)}$	1.69 (0.14)			

Table 6: Performance of each method for high dimensional models $\mathcal{M}_5$–$\mathcal{M}_8$ with $\sigma = 5$ (values for $\mathcal{R}_n(\hat{\alpha}_n)$, $u_n^+(\hat{\alpha}_n)$, $u_n^-(\hat{\alpha}_n)$ in $\mathcal{M}_7$ were rescaled by 10)

	$n = 100$		$n = 500$		$n = 1000$	
	BC	BC-Dat	BC	BC-Dat	BC	BC-Dat
$\mathcal{M}_1$	34.11 (0.61)	35.49 (0.72)	6.20 (0.03)	6.87 (0.06)	3.07 (0.01)	3.49 (0.04)
$\mathcal{M}_2$	63.14 (0.74)	61.14 (0.59)	16.24 (0.16)	15.98 (0.15)	9.26 (0.09)	8.87 (0.08)
$\mathcal{M}_3$	31.39 (0.11)	31.42 (0.11)	6.26 (0.03)	6.81 (0.06)	3.07 (0.01)	3.57 (0.04)
$\mathcal{M}_4$	11.60 (0.05)	11.71 (0.06)	2.13 (0.01)	2.15 (0.02)	1.21 (0.01)	1.24 (0.01)

Table 7: Performance of BC and BC-Dat for variable selection in models $\mathcal{M}_1$-$\mathcal{M}_4$, and their performance for high dimensional models $\mathcal{M}_5$-$\mathcal{M}_8$ is equivalent (which is therefore omitted)

C. Proof of Proposition 1

We first prove identity (14). It is clear that $\underline{\alpha}_n^c$ minimizes $\mathcal{R}_n(\alpha)$ over $\mathcal{A}_n^c$. For all sufficiently large n, condition (9) implies $\mathcal{A}_n^R \subseteq \mathcal{A}_n^c$, which further implies that $\alpha_n^R = \underline{\alpha}_n^c$ and it is the unique element in $\mathcal{A}_n^R$. Based on the argument in (Li, 1987, p.970), condition (13) implies that

$$\max_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \frac{|\sigma^2 d_n(\alpha) - \boldsymbol{e}_n^{\mathrm{T}} H_n(\alpha) \boldsymbol{e}_n|}{n \mathcal{L}_n(\alpha)} \to_p 0. \tag{28}$$

It then follows from

$$\frac{|\mathcal{R}_n(\alpha) - \mathcal{L}_n(\alpha)|}{\mathcal{L}_n(\alpha)} = \frac{|\sigma^2 d_n(\alpha) - \boldsymbol{e}_n^{\mathrm{T}} P_n(\alpha) \boldsymbol{e}_n|}{n \mathcal{L}_n(\alpha)}$$

that

$$\mathcal{R}_n(\alpha) = \mathcal{L}_n(\alpha)\{1 + o_p(1)\} \tag{29}$$

uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. In other words, $\max_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} |\mathcal{R}_n(\alpha)/\mathcal{L}_n(\alpha) - 1| \to_p 0$. For any fixed constant $\varepsilon > 0$, Markov's inequality gives

$$\mathbb{P}\{\mathcal{L}_n(\underline{\alpha}_n^c) \le 2\varepsilon^{-1} n^{-1}\, d_n(\alpha)\sigma^2\} \ge 1 - \frac{\varepsilon}{2}. \tag{30}$$

It follows from (29), (30) that

$$\frac{\mathcal{L}_n(\underline{\alpha}_n^c)}{\mathcal{L}_n(\alpha)} \le \frac{2\varepsilon^{-1} n^{-1}\, d_n(\alpha)\sigma^2}{\mathcal{R}_n(\alpha)\{1 + o_p(1)\}} \le \frac{4\varepsilon^{-1} \mathcal{R}_n(\underline{\alpha}_n^c)}{\mathcal{R}_n(\alpha)} \tag{31}$$

with probability at least $1 - \varepsilon$ for all sufficiently large n. We further conclude from (9) that

$$\mathbb{P}\left\{\mathcal{L}_n(\underline{\alpha}_n^c) < c \min_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \mathcal{L}_n(\alpha)\right\} \to 1 \tag{32}$$

for any fixed constant $0 < c < 1$. Moreover, due to (8), $\underline{\alpha}_n^c$ minimizes $\mathcal{L}_n(\alpha)$ over $\mathcal{A}_n^c$ almost surely. Thus, $\mathbb{P}\{\mathcal{A}_n^L = \{\alpha_n^L\}, \alpha_n^L = \underline{\alpha}_n^c\} \to 1$.

Next, we prove the equivalence of four concepts. From (14), $\mathcal{L}$-consistency and $\mathcal{R}$-consistency are equivalent. Since $\mathcal{R}$-consistency (resp. $\mathcal{L}$-consistency) implies asymptotic risk (resp. loss) efficiency, it remains to prove that asymptotic risk (resp. loss) efficiency implies risk (resp. loss) consistency.

Because of assumption (9), asymptotic risk efficiency implies $\mathbb{P}\{\hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$. Considering the events within $\hat{\alpha}_n \in \mathcal{A}_n^c$, $\mathcal{R}_n(\hat{\alpha}_n)/\mathcal{R}_n(\underline{\alpha}_n^c) \to_p 1$ becomes

$$\frac{d_n(\hat{\alpha}_n)}{d_n(\underline{\alpha}_n^c)} \to_p 1. \tag{33}$$

Because $d_n(\underline{\alpha}_n^c)$ is upper bounded by a fixed constant (denoted by c_0),

$$\mathbb{P}\{\hat{\alpha}_n \neq \underline{\alpha}_n^c\} \leq \mathbb{P}\left(\left|\frac{d_n(\hat{\alpha}_n)}{d_n(\underline{\alpha}_n^c)} - 1\right| \geq c_0^{-1}\right), \tag{34}$$

and thus (33) implies $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c\} \to 1$. It follows from (14) that $\mathbb{P}\{\hat{\alpha}_n = \alpha_n^R\} \to 1$.

Due to (32), asymptotic loss efficiency implies $\mathbb{P}\{\hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$. From (14), it remains to prove that $\hat{\alpha}_n \not\to_p \underline{\alpha}_n^c$ implies $\mathcal{L}_n(\hat{\alpha}_n)/\mathcal{L}_n(\underline{\alpha}_n^c) \not\to_p 1$. In fact, if $\hat{\alpha}_n \not\to_p \underline{\alpha}_n^c$, then it follows from (8), (15), and the boundness of $d_n(\underline{\alpha}_n^c)$ that

$$\frac{\mathcal{L}_n(\hat{\alpha}_n)}{\mathcal{L}_n(\underline{\alpha}_n^c)} - 1 = \sum_{\alpha \in \mathcal{A}_n^c \setminus \{\underline{\alpha}_n^c\}} \frac{\|\boldsymbol{e}_n\|^2_{P_n(\alpha) - P_n(\underline{\alpha}_n^c)}}{\|\boldsymbol{e}_n\|^2_{P_n(\underline{\alpha}_n^c)}} \mathbf{1}_{\hat{\alpha}_n = \alpha} \not\to_p 0. \tag{35}$$

D. Proof of Proposition 2

Case 1). Recall Proposition 1 that $\alpha_n^L = \underline{\alpha}_n^c = \alpha_n^R$ with probability tending to one. To prove that GIC_{λ_n} is $\mathcal{L}$-consistent and $\mathcal{R}$-consistent, it suffices to prove that $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c\} \to 1$. It has been proved under condition (13) that (Shao, 1997, eq.(3.7))

$$G_{n,\lambda_n}(\alpha) = \begin{cases} \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \dfrac{\lambda_n \hat{\sigma}_n^2 d_n(\alpha)}{n} - \dfrac{\boldsymbol{e}_n^{\mathrm{T}} P_n(\alpha) \boldsymbol{e}_n}{n} & \text{if } \alpha \in \mathcal{A}_n^c \\ \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} + \dfrac{(\lambda_n \hat{\sigma}_n^2 - 2\sigma^2) d_n(\alpha)}{n} & \text{if } \alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c \end{cases} \tag{36}$$

where the o_p is taken uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$.

It follows from $\lambda_n \to \infty$ and $\hat{\sigma}_n^2 \to_p \sigma^2$ that $\mathbb{P}\{\lambda_n \hat{\sigma}_n^2 - 2\sigma^2 > 0\} \to 1$; Moreover, from identities (29) and (36), it holds with probability tending to one that

$$\frac{G_{n,\lambda_n}(\underline{\alpha}_n^c) - \|\boldsymbol{e}_n\|^2/n}{G_{n,\lambda_n}(\alpha) - \|\boldsymbol{e}_n\|^2/n} \leq \frac{\lambda_n \mathcal{R}_n(\underline{\alpha}_n^c)}{\mathcal{R}_n(\alpha)} \{1 + o_p(1)\}$$

uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. It then follows from condition (16) that

$$\mathbb{P}\left\{ \max_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} \frac{G_{n,\lambda_n}(\underline{\alpha}_n^c) - \|\boldsymbol{e}_n\|^2/n}{G_{n,\lambda_n}(\alpha) - \|\boldsymbol{e}_n\|^2/n} < q \right\} \to 1 \tag{37}$$

for some constant $q \in (0, 1)$. Therefore,

$$\mathbb{P}\left\{ \min_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c} G_{n,\lambda_n}(\alpha) > G_{n,\lambda_n}(\underline{\alpha}_n^c) \right\} \to 1,$$

which implies $\mathbb{P}\{\hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$. It remains to prove that $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c,\ \hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$, or equivalently,

$$\mathbb{P}\left\{ \min_{\alpha \in \mathcal{A}_n^c \setminus \{\underline{\alpha}_n^c\}} G_{n,\lambda_n}(\alpha) > G_{n,\lambda_n}(\underline{\alpha}_n^c) \right\} \to 1. \tag{38}$$

For any $\alpha \in \mathcal{A}_n^c$,

$$\begin{aligned} & G_{n,\lambda_n}(\alpha) - G_{n,\lambda_n}(\underline{\alpha}_n^c) \\ & = \frac{1}{n}\hat{\sigma}_n^2\{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}\biggl(\lambda_n - \frac{\|\boldsymbol{e}_n\|^2_{P_n(\alpha)-P_n(\underline{\alpha}_n^c)}}{\hat{\sigma}_n^2\{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}}\biggr). \end{aligned} \tag{39}$$

From (Whittle, 1960, Thm.2),

$$E\biggl(\frac{\|\boldsymbol{e}\|^2_{P_n(\alpha)-P_n(\underline{\alpha}_n^c)}}{d_n(\alpha) - d_n(\underline{\alpha}_n^c)} - \sigma^2\biggr)^{2m_2} \leq C\{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_2}. \tag{40}$$

By Markov's inequality, we obtain

$$\mathbb{P}\biggl\{\frac{\|\boldsymbol{e}\|^2_{P_n(\alpha)-P_n(\underline{\alpha}_n^c)}}{d_n(\alpha) - d_n(\underline{\alpha}_n^c)} - \sigma^2 > \delta\biggr\} \leq C\delta^{-2m_2}\{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_2} \tag{41}$$

for any constant $\delta > 0$. This together with (17) implies that

$$\max_{\alpha \in \mathcal{A}_n^c, \alpha \neq \underline{\alpha}_n^c} \frac{\|\boldsymbol{e}\|^2_{P_n(\alpha)-\mathbb{P}_n(\underline{\alpha}_n^c)}}{d_n(\alpha) - d_n(\underline{\alpha}_n^c)} = O_p(1). \tag{42}$$

Applying condition (42) and $\lambda_n \to \infty$ to identity (39), we conclude that (38) holds.

Case 2). It has been proved under conditions $\hat{\sigma}_n^2 \to_p \sigma^2$ and (13) that (Shao, 1997, eq.(3.2))

$$G_{n,2}(\alpha) = \begin{cases} \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \dfrac{2\hat{\sigma}_n^2 d_n(\alpha)}{n} - \dfrac{\boldsymbol{e}_n^{\mathrm{T}} P_n(\alpha)\boldsymbol{e}_n}{n} & \text{if } \alpha \in \mathcal{A}_n^c \\ \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} & \text{if } \alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c \end{cases} \tag{43}$$

where the o_p is taken uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$ (as before). If $\mathrm{card}(\mathcal{A}_n^c) = \emptyset$, it is clear that minimizing $G_{n,2}(\alpha)$ guarantees asymptotic loss efficiency. If $\mathrm{card}(\mathcal{A}_n^c) = 1$, then conditions (9), (29), and (43) imply both $\mathbb{P}\{G_{n,2}(\alpha) > G_{n,2}(\underline{\alpha}_n^c)\} \to 1$ and $\mathcal{L}_n(\underline{\alpha}_n^c) = o_p\{\mathcal{L}_n(\alpha)\}$ uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. Thus, the selected model is asymptotically loss efficient. The asymptotic risk efficiency can be similarly proved using (29).

E. Proof of Theorem 1

A more general result is summarized in Proposition 4. To prove (21), we only need to verify the assumptions there. Assumption ii implies assumption ii in Proposition 4. Assumption iv trivially implies assumption iv(c) in Proposition 4. Assumptions v and (18) and the nested model imply assumption iv(e) in Proposition 4. Assumption (20) and the model being nested imply assumption iv(f)(g) in Proposition 4.

Moreover, in the particular case of $\Delta_n(\alpha) = c\, d(\alpha)^{-\gamma}$, it is sufficient to simply choose $n' = n \cdot (1 + n^{-\zeta})$ with any $\zeta \in (\frac{1}{2}, \frac{1}{\gamma+1})$.

Proposition 4. *Assume that the noises $e_1, \ldots, e_n$ are i.i.d. Gaussian with zero mean and variance σ^2. Assume that $\mathcal{A}_n^c = \emptyset$. Suppose that for all sufficiently large n*

(i) $\Delta_n(\bar{\alpha}_n) \to 0$, $d_n/n \to 0$;

(ii) Condition (13) holds;

(iii) *card*$(\mathcal{A}_n^R) < C$ *for some fixed constant* C;

(iv) There exist n' ($n' > n$) that is a function of n and satisfies

(a) $n^{-1/2}(n - n') \to 0$,

(b) $n^{-1/2}\{d(\alpha_n^R) - d(\alpha_{n'}^R)\} \to 0$,

(c) $n^{1/2}\{\Delta_n(\alpha_n^R) - \Delta_{n'}(\alpha_n^R)\} \to 0$,

(d) $n\{\Delta_n(\alpha_n^R) - \Delta_n(\alpha_{n'}^R)\} \to \infty$,

(e)

$$\limsup_{n\to\infty} \frac{n\{\Delta_n(\alpha_n^R) - \Delta_n(\alpha_{n'}^R)\} - n'\{\Delta_{n'}(\alpha_n^R) - \Delta_{n'}(\alpha_{n'}^R)\}}{\sqrt{n\{\Delta_n(\alpha_n^R) - \Delta_n(\alpha_{n'}^R)\}}} < 2\sqrt{e},$$

(f) $\mathcal{A}_n^R \cap \mathcal{A}_{n'}^R = \emptyset$,

(g) $\mathcal{A}_n^R \in \mathcal{A}_{n'}$*,* $\mathcal{A}_{n'}^R \in \mathcal{A}_n$*,*

(h) $\alpha_n^R \subsetneq \alpha_{n'}^R$*,*

for any $\alpha_n^R \in \mathcal{A}_n^R$ *and* $\alpha_{n'}^R \in \mathcal{A}_{n'}^R$*.*

Then the conditions of Proposition 2(ii) are satisfied (so that asymptotic efficiency can be achieved), while for any normal selection criterion ψ

$$\limsup_{n\to\infty} \mathbb{P}_n\{\psi_n(\boldsymbol{y}_n, \mathcal{A}_n) \notin \mathcal{A}_n^R\} > 0, \tag{44}$$

where $\mathbb{P}_n$ *denotes the probability under the distribution of* $\boldsymbol{y}_n$*.*

Proof of Proposition 4:

It can be verified that the conditions of Proposition 2(ii) are met. In particular, assumption (i) guarantees $\hat{\sigma}_n^2 \to_p \sigma^2$ (by Remark 1).

For notational convenience, let $n_1 = n$ and $n_2 = n'$ (which depends only on n_1). Since

$$\limsup_{n\to\infty} \mathbb{P}_n\{\psi_n(\boldsymbol{y}_n, \mathcal{A}_n) \notin \mathcal{A}_n^R\} \geq \liminf_{n_1\to\infty}\left(\frac{1}{2}\sum_{k=1,2} \mathbb{P}_{n_k}\{\psi_{n_k}(\boldsymbol{y}_{n_k}, \mathcal{A}_{n_k}) \notin \mathcal{A}_{n_k}^R\right),$$

and property (P1) implies that $\psi_{n_k}(\boldsymbol{y}_{n_k}, \mathcal{A}_{n_k}) = \phi_{n_k}(\boldsymbol{s}_{n_k})$, we only need to prove that

$$\liminf_{n_1\to\infty}\left(\frac{1}{2}\sum_{k=1,2} \mathbb{P}_{n_k}\{\phi_{n_k}(\boldsymbol{s}_{n_k}) \notin \mathcal{A}_{n_k}^R\right) > 0.$$

Using property (P2) and assumption (iii), it suffices to prove that for any $\alpha_{n_1}^R \in \mathcal{A}_{n_1}$, $\alpha_{n_2}^R \in \mathcal{A}_{n_2}$,

$$\sum_{k=1,2} \mathbb{P}_{n_k}\{\phi_{n_k}(\boldsymbol{S}) \notin \mathcal{A}_{n_k}^R\} \geq \delta \tag{45}$$

for some fixed constant $\delta > 0$, where

$$\boldsymbol{S} \triangleq [S_{n_k}(\alpha_{n_1}^R), S_{n_k}(\alpha_{n_2}^R)]^{\mathrm{T}}$$

under probability $\mathbb{P}_{n_k}$. Note that the distribution of $\boldsymbol{S}$ depends on whether $k = 1, 2$.

To prove (45), we shall use an adaptation of Le Cam's method. For each $\alpha \in \mathcal{A}_{n_k}$, we let $\mathbf{1}_{k,\alpha}$ denote the card$(\mathcal{A}_{n_k}) \times 1$ zero-one vector that contains a unique '1' representing α, and let $I_{\mathcal{A}_{n_k}^R} = \{\mathbf{1}_\alpha : \alpha \in \mathcal{A}_{n_k}^R\}$ for $k = 1, 2$. For a point x and set S in the Euclidean space, we define $D(x, S) = \inf_{s \in S} \|x - s\|$. For any selection criterion ϕ_{n_k} restricted to $\boldsymbol{S}$, it defines the test statistic

$$T(\boldsymbol{S}) = \begin{cases} 1 & \text{if } D(\mathbf{1}_{\phi_{n_1}(\boldsymbol{S})}, I_{\alpha_{n_1}^R}) \leq D(\mathbf{1}_{\phi_{n_2}(\boldsymbol{S})}, I_{\alpha_{n_2}^R}) \\ 2 & \text{otherwise.} \end{cases}$$

If $T(\boldsymbol{S}) = 1$, then assumption iv(f) guarantees that $D(\mathbf{1}_{\phi_{n_2}(\boldsymbol{S})}, I_{\mathcal{A}_{n_2}^R}) \neq 0$, which further implies that $D(\mathbf{1}_{\phi_{n_2}(\boldsymbol{S})}, I_{\mathcal{A}_{n_2}^R}) = \sqrt{2}$. Therefore,

$$\begin{aligned} \mathbb{P}_{n_2}\big\{\phi_{n_2}(\boldsymbol{S}) \notin \mathcal{A}_{n_2}^R\big\} &= \frac{1}{\sqrt{2}} E_2\Big\{D(\mathbf{1}_{\phi_{n_2}(\boldsymbol{S})}, I_{\mathcal{A}_{n_2}^R})\Big\} \\ &\geq \frac{1}{\sqrt{2}} E_2\Big\{D(\mathbf{1}_{\phi_{n_2}(\boldsymbol{S})}, I_{\mathcal{A}_{n_2}^R}) 1_{T(\boldsymbol{S})=1}\Big\} = \mathbb{P}_{n_2}\{T(\boldsymbol{S}) = 1\} \end{aligned}$$

where $1_{(\cdot)}$ is the indicator function, and E_k is the expectation with respect to $\mathbb{P}_{n_k}$ $(k = 1, 2)$. Similarly, $\mathbb{P}_{n_1}\big\{\phi_{n_1}(\boldsymbol{S}) \notin \mathcal{A}_{n_1}^R\big\} \geq \mathbb{P}_{n_1}\{T(\boldsymbol{S}) = 2\}$. Thus, we obtain

$$\sum_{k=1,2} \mathbb{P}_{n_k}\big\{\phi_{n_k}(\boldsymbol{S}) \notin \mathcal{A}_n^{R,k}\big\} \geq \sum_{k=1,2} \mathbb{P}_{n_k}\{T(\boldsymbol{S}) \neq k\}. \tag{46}$$

Let $\mathbb{P}_k$ and p_k denote the probability distribution of $\boldsymbol{S}$ and its density function under sample size n_k $(k = 1, 2)$, respectively. The right hand side of (46) is minimized by the Neyman-Pearson test

$$T_{\text{NP}}(\boldsymbol{S}) = \begin{cases} 1 & \text{if } p_1(\boldsymbol{S}) > p_2(\boldsymbol{S}) \\ 2 & \text{otherwise} \end{cases}.$$

Direct calculation gives

$$\sum_{k=1,2} \mathbb{P}_{n_k}\{T(\boldsymbol{S}) \neq k\} \geq \sum_{k=1,2} \mathbb{P}_{n_k}\{T_{\text{NP}}(\boldsymbol{S}) \neq k\} = 1 - D_{\text{TV}}(\mathbb{P}_1, \mathbb{P}_2). \tag{47}$$

Combining (45), (46), and (47), it remains to prove that

$$\limsup_{n\to\infty} D_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) < 1.$$

For notational convenience, let

$$\begin{aligned} U_1 &\triangleq S_{n_1}(\alpha^R_{n_1}), \quad U_2 \triangleq S_{n_1}(\alpha^R_{n_1}) - S_{n_1}(\alpha^R_{n_2}) \\ V_1 &\triangleq S_{n_2}(\alpha^R_{n_1}), \quad V_2 \triangleq S_{n_2}(\alpha^R_{n_1}) - S_{n_2}(\alpha^R_{n_2}) \end{aligned}$$

Let d_1, d_2 denote the dimensions of $\alpha^R_{n_1}, \alpha^R_{n_2}$, respectively. Direct calculation shows that U_1, U_2, V_1, V_2 follow non-central chi-squared (nc-χ^2) distributions with non-centrality parameters and degrees of freedom given below

$$\begin{aligned} &U_1 \sim \text{nc-}\chi^2(s^U_1, r^U_1), \quad s^U_1 = n_1 \Delta_{n_1}(\alpha^R_{n_1}), \quad r^U_1 = n_1 - d_1, \\ &V_1 \sim \text{nc-}\chi^2(s^V_1, r^V_1), \quad s^V_1 = n_2 \Delta_{n_2}(\alpha^R_{n_1}), \quad r^V_1 = n_2 - d_2, \\ &U_2 \sim \text{nc-}\chi^2(s^U_2, r^U_2), \quad s^U_2 = n_1 \{\Delta_{n_1}(\alpha^R_{n_1}) - \Delta_{n_1}(\alpha^R_{n_2})\}, \quad r^U_2 = d_2 - d_1, \\ &V_2 \sim \text{nc-}\chi^2(s^V_2, r^V_2), \quad s^V_2 = n_2 \{\Delta_{n_2}(\alpha^R_{n_1}) - \Delta_{n_2}(\alpha^R_{n_2})\}, \quad r^V_2 = d_2 - d_1, \end{aligned}$$

It follows from the assumptions of Gaussian noise and assumption iv(h) that U_1, U_2 are independent, and V_1, V_2 are independent. As a result,

$$D_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) = D_{\mathrm{TV}}(\mathbb{P}_{[U_1, U_1 - U_2]}, \mathbb{P}_{[V_1, V_1 - V_2]}) = D_{\mathrm{TV}}(\mathbb{P}_{[U_1, U_2]}, \mathbb{P}_{[V_1, V_2]}).$$

From Lemma 1,

$$D_{\mathrm{TV}}(\mathbb{P}_U, \mathbb{P}_V) \leq \sum_{k=1}^{2} D_{\mathrm{TV}}(\mathbb{P}_{U_k}, \mathbb{P}_{V_k}).$$

Next we prove that

$$\lim_{n\to\infty} D_{\mathrm{TV}}(\mathbb{P}_{U_1}, \mathbb{P}_{V_1}) = 0, \tag{48}$$

$$\limsup_{n\to\infty} D_{\mathrm{TV}}(\mathbb{P}_{U_2}, \mathbb{P}_{V_2}) < 1. \tag{49}$$

Let Z denote the standard Gaussian random variable. By triangle inequality and Lemmas 2 we have

$$\begin{aligned}
D_{\mathrm{TV}}(\mathbb{P}_{U_1}, \mathbb{P}_{V_1}) &= D_{\mathrm{TV}}(\mathbb{P}_{(U_1 - r_1^U - s_1^U)/\sqrt{n_1}}, \mathbb{P}_{(V_1 - r_1^U - s_1^U)/\sqrt{n_1}}) \\
&\leq C_1 + D_{\mathrm{TV}}(\mathbb{P}_Z, \mathbb{P}_{(V_1 - r_1^U - s_1^U)/\sqrt{n_1}}) \\
&\leq C_1 + D_{\mathrm{TV}}(\mathbb{P}_{aZ+b}, \mathbb{P}_{(V_1 - r_1^V - s_1^V)/\sqrt{n_2}}) \\
&\leq C_1 + C_2 + C_3
\end{aligned}$$

where

$$\begin{aligned}
&C_1 \triangleq D_{\mathrm{TV}}(\mathbb{P}_{(U_1 - r_1^U - s_1^U)/\sqrt{n_1}}, \mathbb{P}_Z) \\
&C_2 \triangleq D_{\mathrm{TV}}(\mathbb{P}_{aZ+b}, \mathbb{P}_Z), \quad C_3 \triangleq D_{\mathrm{TV}}(\mathbb{P}_Z, \mathbb{P}_{(V_1 - r_1^V - s_1^V)/\sqrt{n_2}}), \\
&a \triangleq \sqrt{\frac{n_1}{n_2}}, \quad b \triangleq \frac{(r_1^U - r_1^V) + (s_1^U - s_1^V)}{\sqrt{n_2}}.
\end{aligned}$$

From Lemma 4 and Lemma 2, (48) holds as long as the following conditions hold.

$$\frac{n_1}{n_2} = 1 + o(1) \tag{50}$$

$$\frac{n_1 \Delta_{n_1}(\alpha_{n_1}^R) - n_2 \Delta_{n_2}(\alpha_{n_1}^R)}{\sqrt{n_1}} = o(1) \tag{51}$$

$$\frac{(n_1 - d_1) - (n_2 - d_2)}{\sqrt{n_1}} = o(1) \tag{52}$$

as $n = n_1 \to \infty$. In fact, assumption iv(a) implies (50), assumptions ii, iv(a)-(c) imply (51), and assumptions iv(a)(b) imply (52).

To prove (49), by Lemma 6, it suffices to show that

$$s_2^U \to \infty, \qquad \limsup_{n\to\infty} \frac{s_2^U - s_2^V}{\sqrt{s_2^U}} < 2\sqrt{e}. \tag{53}$$

In fact, the two inequalities in (53) are implied by assumptions iv(d)&(e), respectively.

F. Technical lemmas for the proof of Theorem 1 and Proposition 4

We first introduce some new notation. Total variation distance of two probability measures P, Q on a sigma-algebra $\mathcal{F}$ of subsets of the sample space $\mathcal{X}$ is defined by $D_{\text{TV}}(P,Q) = \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$. It is related to the L_1 distance by the identity $2D_{\text{TV}}(P,Q) = \|f_P - f_Q\|_1$ where f_P, f_Q are absolutely continuous probability density functions.

Lemma 1. *Suppose that $U = [U_1, \ldots, U_k]^{\text{T}}$ and $V = [V_1, \ldots, V_k]^{\text{T}}$ are random variables defined on a common probability space such that all U_i's are independent and all V_i's are independent. Then*

$$D_{\text{TV}}(\mathbb{P}_U, \mathbb{P}_V) \leq \sum_{i=1}^{k} D_{\text{TV}}(\mathbb{P}_{U_i}, \mathbb{P}_{V_i}).$$

Proof. The proof follows from the definition of total variation distance and the elementary inequality $|\prod_{i=1}^k a_i - \prod_{i=1}^k b_i| \leq \sum_{i=1}^k |a_i - b_i|$ for all real values a_i's and b_i's that $|a_i| \leq 1$, $|b_i| \leq 1$.

Lemma 2. *Suppose that U and V are random variables, and a, b are deterministic values. Then*

$$D_{\text{TV}}(\mathbb{P}_{aU+b}, \mathbb{P}_{aV+b}) = D_{\text{TV}}(\mathbb{P}_U, \mathbb{P}_V).$$

Proof. The scale and shift invariance directly follows from the definition of total variation distance.

Lemma 3. *Suppose that U is a univariate Gaussian random variable, and a_n, b_n are deterministic sequences such that $a_n \to 1$ and $b_n \to 0$ as n goes to infinity. Then*

$$\lim_{n\to\infty} D_{\text{TV}}(\mathbb{P}_{a_n U + b_n}, \mathbb{P}_U) = 0.$$

Moreover, double indexed sequences $a_{n,i}$ and $b_{n,i}$ satisfy

$$\max_{i\in I_n} |a_{n,i} - 1| \to 0, \quad \max_{i\in I_n} |b_{n,i}| \to 0, \tag{54}$$

for some set $I_n \in \{1, \ldots, n\}$, then

$$\lim_{n\to\infty} \max_{i \in I_n} D_{\mathrm{TV}}(\mathbb{P}_{a_{n,i}U+b_{n,i}}, \mathbb{P}_U) = 0.$$

The result for the case of $a_{n,i}$ and $b_{n,i}$ can be similarly proved.

Proof. Suppose that $U \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Let $\mu_1 = a_n\mu_0 + b_n, \sigma_1^2 = a_n^2\sigma_0^2$. The Kullback-Leibler divergence from $\mathbb{P}_{a_nU+b_n}$ to $\mathbb{P}_U$ is

$$\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + \frac{(\mu_1-\mu_0)^2}{\sigma_1^2} - 1 + \log\frac{\sigma_1^2}{\sigma_0^2}\right),$$

which goes to zero as $n \to \infty$. The desired result then follows from Pinsker's inequality.

The following lemma shows that with large degrees of freedom (compared with non-centrality parameters), standardized noncentral chi-squared distributions converge to standard Gaussian distribution in total variation distance (which is stronger than the usual Kolmogorov distance).

Lemma 4. *Suppose that X_n is a sequence of noncentral chi-squared random variables with degrees of freedom r_n and non-centrality parameters s_n. Suppose that r_n, s_n, t_n are deterministic sequences such that r_n's are positive integers and as $n \to \infty$,*

$$s_n \to \infty, \qquad \frac{s_n}{r_n} \to 0, \tag{55}$$

$$\frac{r_n}{t_n} \to 1. \tag{56}$$

Then $D_{\mathrm{TV}}(\mathbb{P}_{\tilde{X}_n}, \mathbb{P}_Z) \to 0$ as $n \to \infty$, where

$$\tilde{X}_n = \frac{X_n - (r_n + s_n)}{\sqrt{t_n}},$$

and $\mathbb{P}_Z$ denotes a standard Gaussian distribution.

Proof. For any fixed $\delta > 0$, we prove that

$$D_{\mathrm{TV}}(p_{\tilde{X}_n}, p_Z) \le \delta \tag{57}$$

for all sufficiently large n. The probability density function of X_n is given by

$$f_{X_n}(x; r_n, s_n) = \sum_{i=0}^{\infty} h_{s_n/2}(i) f_{Y_{r_n+2i}}(x), \tag{58}$$

where $h_{s_n/2}(i)$ is the distribution function of Poisson with mean $s_n/2$, and Y_r is distributed as chi-squared with r degrees of freedom. Condition (55) and the Chebyshev's inequality imply that

$$\lim_{n\to\infty} \sum_{i\in I_n} h_{s_n/2}(i) = 1, \text{ where } I_n \triangleq \left\{ i : \left|i - \frac{s_n}{2}\right| \le (s_n r_n)^{1/4} \right\}.$$

By triangle inequality and the definition of $h_{s_n/2}(\cdot)$,

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_{\tilde{X}_n}, \mathbb{P}_Z) &\le \frac{1}{2} \sum_{i=0}^{\infty} h_{s_n/2}(i) \left\| f_{\{Y_{r_n+2i}-(r_n+s_n)\}/\sqrt{t_n}} - f_Z \right\|_1 \\ &\le \frac{\delta}{3} + \frac{1}{2} \sum_{i\in I_n} h_{s_n/2}(i) \left\| f_{\{Y_{r_n+2i}-(r_n+s_n)\}/\sqrt{t_n}} - f_Z \right\|_1 \\ &\le \frac{\delta}{3} + \frac{1}{2} \max_{i\in I_n} \left\| f_{\{Y_{r_n+2i}-(r_n+s_n)\}/\sqrt{t_n}} - f_Z \right\|_1 \end{aligned} \tag{59}$$

for all sufficiently large n. By Lemma 2,

$$\begin{aligned} &\left\| f_{\{Y_{r_n+2i}-(r_n+s_n)\}/\sqrt{t_n}} - f_Z \right\|_1 = \left\| f_{\{Y_{r_n+2i}-(r_n+2i)\}/\sqrt{r_n+2i}} - f_{a_n Z+b_n} \right\|_1 \\ &\le \left\| f_{\{Y_{r_n+2i}-(r_n+2i)\}/\sqrt{r_n+2i}} - f_Z \right\|_1 + \left\| f_{a_n Z+b_n} - f_Z \right\|_1, \end{aligned} \tag{60}$$

where

$$a_n \triangleq \sqrt{\frac{t_n}{r_n+2i}}, \quad b_n \triangleq \frac{s_n - 2i}{\sqrt{r_n+2i}}. \tag{61}$$

It can be proved that condition (54) holds. Then Lemma 3 implies that

$$\| f_{a_n Z+b_n} - f_Z \|_1 \le \frac{2\delta}{3} \tag{62}$$

for all sufficiently large n. Moreover, by (Prohorov, 1952; Bally et al., 2016, Theorem 2.2)

$$\left\| f_{\{Y_{r_n+2i}-(r_n+2i)\}/\sqrt{r_n+2i}} - f_Z \right\|_1 \le \frac{2\delta}{3} \tag{63}$$

for all $i \in I_n$ for all sufficiently large n.

The proof is completed by combining inequalities (59), (60), (62), and (63).

The following is a technical lemma stating that a Poisson distribution has negligible total variation distance to its shift as long as such shift is small.

LEMMA 5. *Suppose that X_n is a sequence of Poisson random variables with mean $\tau_n \to \infty$, and ρ_n is a deterministic sequence of integers such that $\rho_n = o(\sqrt{\tau_n})$. Then $D_{\text{TV}}(\mathbb{P}_{X_n}, \mathbb{P}_{X_n - \rho_n}) \to 0$ as $n \to \infty$.*

PROOF. It follows from the assumptions and the Chebyshev's inequality that

$$\lim_{n\to\infty} \sum_{i\in I_n} h_{\tau_n}(i) = 1, \text{ where } I_n \triangleq \left\{ i : \left| i - \tau_n \right| \le (\tau_n \rho_n)^{1/4} \right\}.$$

For any fixed $\delta > 0$, we have

$$\sum_{i \notin I_n} h_{\tau_n}(i) \le \frac{\delta}{3}, \qquad \sum_{i-\rho_n \notin I_n} h_{\tau_n}(i) \le \frac{\delta}{3}, \tag{64}$$

for all sufficiently large n. For all $i \in I_n$ we have $i + \tau_n \sim i$. By Stirling's formula,

$$\begin{aligned} \frac{\mathbb{P}(X_n = i)}{\mathbb{P}(X_n - \rho_n = i)} &= \frac{\tau_n^{-\rho_n} (i + \tau_n)!}{i!} \\ &\sim \tau_n^{-\rho_n} \left(\frac{i + \rho_n}{i} \right)^{1/2} \left(1 + \frac{\rho_n}{i} \right)^{i} \left(\frac{i + \rho_n}{e} \right)^{\rho_n} \\ &\sim \tau_n^{-\rho_n} (i + \rho_n)^{\rho_n} = o(1) \end{aligned} \tag{65}$$

uniformly for $i \in I_n$ as $n \to \infty$. Combining (64) and (65) we obtain

$$D_{\text{TV}}(\mathbb{P}_{X_n}, \mathbb{P}_{X_n - \rho_n}) \le \frac{2\delta}{3} + \frac{1}{2} \sum_{i \in I_n} |\mathbb{P}(X_n = i) - \mathbb{P}(X_n - \rho_n = i)| \le \delta$$

for all sufficiently large n.

The following lemma shows that with the total variation distance between two noncentral chi-squared distributions is close to zero as long as their difference in mean is small compared with their standard deviations.

LEMMA 6. *Suppose that $X_{1,n}, X_{2,n}$ are two sequences of noncentral chi-squared random variables with degrees of freedom $r_{1,n}, r_{2,n}$ and noncentrality parameters $s_{1,n}, s_{2,n}$, respectively. Suppose that $r_{k,n}, s_{k,n}$ are deterministic sequences such that $r_{k,n}$'s ($k = 1, 2$) are positive integers and as $n \to \infty$,*

$$s_{1,n} \to \infty, \tag{66}$$

$$\frac{r_{1,n} - r_{2,n}}{\sqrt{s_{1,n}}} \to 0; \tag{67}$$

$$\limsup_{n\to\infty} \frac{s_{1,n} - s_{2,n}}{\sqrt{s_{1,n}}} < 2\sqrt{e}. \tag{68}$$

Then $\limsup_{n\to\infty} D_{\mathrm{TV}}(p_{X_{1,n}}, p_{X_{2,n}}) < 1$, where p_X denotes the distribution of X.

PROOF. Choose any fixed δ such that

$$0 < 3\delta < 1 - \limsup_{n\to\infty} \frac{s_{1,n} - s_{2,n}}{2\sqrt{e \cdot s_{1,n}}}.$$

We prove that $D_{\mathrm{TV}}(\mathbb{P}_{X_{1,n}}, \mathbb{P}_{X_{2,n}}) \le 1 - \delta$ for all sufficiently large n. For simplicity, we assume that $r_{1,n} - r_{2,n}$ are even numbers. Other cases can be similarly proved. Define $I_n = \{i : |i - s_{1,n}/2| < s_{1,n}^{1/2+\epsilon}\}$ for any $\epsilon \in (0, 1/2)$. Similar to the argument in (64) we have

$$\sum_{i \notin I_n} h_{s_{1,n}/2}(i) \le \frac{\delta}{4}, \qquad \sum_{i-(r_{1,n}-r_{2,n})/2 \notin I_n} h_{s_{2,n}/2}(i) \le \frac{\delta}{4}. \tag{69}$$

Using (69) and identity (58) we have

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_{X_{1,n}}, \mathbb{P}_{X_{2,n}}) &\le \frac{\delta}{2} + \sum_{\substack{i \in I_n, \\ j = i + (r_{1,n} - r_{2,n})/2}} |h_{s_{1,n}/2}(i) - h_{s_{2,n}/2}(j)| \cdot \|f_{Y_{r_{1,n}+2i}}(y)\|_1 \\ &\le \frac{\delta}{2} + D_{\mathrm{TV}}(\mathbb{P}_{Y_{1,n}}, \mathbb{P}_{Y_{2,n}-(r_{1,n}-r_{2,n})/2}) \\ &\le \frac{\delta}{2} + D_{\mathrm{TV}}(\mathbb{P}_{Y_{1,n}}, \mathbb{P}_{Y_{2,n}}) + D_{\mathrm{TV}}(\mathbb{P}_{Y_{2,n}}, \mathbb{P}_{Y_{2,n}-(r_{1,n}-r_{2,n})/2}) \end{aligned} \tag{70}$$

where $Y_{1,n}, Y_{2,n}$ are Poisson random variables with means $s_{1,n}/2, s_{2,n}/2$, respectively. Lemma 5 and conditions (68), (67) imply that

$$D_{\mathrm{TV}}(\mathbb{P}_{Y_{2,n}}, \mathbb{P}_{Y_{2,n}-(r_{1,n}-r_{2,n})/2}) \le \frac{\delta}{2} \tag{71}$$

for all sufficiently large n. Using (Adell and Jodrá, 2006, Inequality 2.2),

$$\begin{aligned} D_{\mathrm{TV}}(\mathbb{P}_{Y_{1,n}}, \mathbb{P}_{Y_{2,n}}) &\le \sqrt{\frac{2}{e}} \left| \sqrt{\frac{s_{1,n}}{2}} - \sqrt{\frac{s_{2,n}}{2}} \right| \\ &= \frac{1}{\sqrt{e}} \frac{|s_{1,n} - s_{2,n}|}{\sqrt{s_{1,n}} + \sqrt{s_{2,n}}} \\ &\le 1 - 2\delta \end{aligned} \tag{72}$$

for all sufficiently large n, where the last inequality follows from condition (68). Combining (70)-(72) we complete the proof.

G. Proof of Theorem 2

Remark 2 (Interpretation of each condition). *The conditions that do not appear in Proposition 2 are explained below. In view of the previous intuitive arguments, (23) is a regularity condition to guarantee that GIC$_2$ does not overfit too much, namely $d_n(\hat{\alpha}_{\mathrm{GIC}_2}) = d_n(\underline{\alpha}_n^c) + O_p(1)$ in parametric scenario. Conditions (22) and (24) ensure that (6) are sufficiently large penalties to eliminate all the remaining overfitting candidates. We note that (23) together with (24) gives a stronger version of condition (17). They are trivially satisfied for nested model classes, given that $E(e_1^{4+\delta}) < \infty$ for some constant $\delta > 0$.*

In the nonparametric scenario, (25) controls the divergence rate of λ_n so that the optimal performance at the boundary $\hat{\alpha}_{\mathrm{GIC}_2}$ can be reached. The regularity of $\mathcal{R}_n(\cdot)$ serves to further ensure that $\lambda_n < \tilde{q}\, d_n(\hat{\alpha}_{\mathrm{GIC}_2})/\log d_n(\hat{\alpha}_{\mathrm{GIC}_2})$ for some constant $\tilde{q} < 1$, which is used in the proof.

The regularity of $\mathcal{R}_n(\cdot)$ means that the optimal (sequence of) risks $\mathcal{R}_n(\alpha_n^R)$ can be approached by α_n only if the model dimension $d_n(\alpha_n)$ and one of $\{d_n(\alpha) : \alpha \in \mathcal{A}_n^R\}$ are comparable. It is a mild assumption. Consider, for example, $\mathcal{A}_n = \{\alpha_1, \alpha_2, \ldots\}, \alpha_d = \{1, \ldots, d\}$ with algebraic decaying model approximation error: $\Delta_n(\alpha_d) = c\, d^{-\gamma}$. For this case, we have $d_n(\alpha_n^R) \sim n^r$ $(0 < r < 1)$. In view of Remark 1, suppose that in the parametric case, a fixed $d_n(\underline{\alpha}_n^c)$ and condition (10) are further assumed, then $\lambda_n \sim n^\tau$ with any $0 < \tau < r$ suffices to meet all the conditions (16), (22), (25). In fact, the assumption of regular $\mathcal{R}_n(\cdot)$ is mostly for easy verification when $\mathcal{R}_n(\cdot)$ is known, and it is not essential. Instead, one may require $\lambda_n \leq q\, d_n(\hat{\alpha}_{\mathrm{GIC}_2})/\log d_n(\hat{\alpha}_{\mathrm{GIC}_2})$ directly. The condition $d_n(\alpha_n^R) \geq d_0$ for some constant $d_0 > 1$ is mild since $d_n(\alpha_n^R) \to \infty$ is common in nonparametric model classes (as we have discussed before). It is only to guarantee that (25) implies $\lambda_n \leq q d_n(\alpha)/\log d_n(\alpha)$ for all

$\alpha \in \mathcal{A}_n^R$.

We note that the divergence rates of d_n and card($\mathcal{A}_n$) are not explicitly required in Theorem 2 (or in Proposition 1 and 2). They implicitly play a role in conditions such as (16). Also, $\underline{\alpha}_n^c$ is not required to be bounded by a fixed constant.

Proof of Theorem 2:

Throughout the proof, let $\hat{\alpha}_n$ denote the model selected by Bridge criterion.

$\mathcal{L}$-consistency and $\mathcal{R}$-consistency under $\mathcal{A}_n^c \neq \emptyset$

Step 1): We first prove that $d_n(\hat{\alpha}_{\mathrm{GIC}_2}) - d_n(\underline{\alpha}_n^c)$ is stochastically bounded. Identity (43), $\hat{\sigma}_n^2 \to_p \sigma^2$, and conditions (9), (29) imply that

$$\mathbb{P}\biggl\{ \min_{\alpha \in \mathcal{A}_n \backslash \mathcal{A}_n^c} G_{n,2}(\alpha) > G_{n,2}(\underline{\alpha}_n^c) \biggr\} \to 1.$$

Thus, it suffices to prove that the dimension of $\arg\min_{\alpha \in \mathcal{A}_n^c} G_{n,2}(\alpha)$ is $d_n(\underline{\alpha}_n^c) + O_p(1)$. For $\alpha \in \mathcal{A}_n^c$, from (43), the event $G_{n,2}(\alpha) < G_{n,2}(\underline{\alpha}_n^c)$ implies that

$$2\hat{\sigma}_n^2 \{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\} < \|\boldsymbol{e}\|_{P_n(\alpha)}^2 - \|\boldsymbol{e}\|_{\mathbb{P}_n(\underline{\alpha}_n^c)}^2,$$

or by (8),

$$\frac{\|\boldsymbol{e}\|^2_{P_n(\alpha) - \mathbb{P}_n(\underline{\alpha}_n^c)}}{d_n(\alpha) - d_n(\underline{\alpha}_n^c)} - \sigma^2 > 2\hat{\sigma}_n^2 - \sigma^2. \tag{73}$$

For any fixed $\varepsilon \in (0,1)$, since $\hat{\sigma}_n \to_p \sigma^2$, there exists $n_1 \in \mathbb{N}$ such that for all $n > n_1$,

$$\mathbb{P}\biggl\{ 2\hat{\sigma}_n^2 - \sigma^2 > \frac{1}{2}\sigma^2 \biggr\} > 1 - \frac{\varepsilon}{2}. \tag{74}$$

Similar to the proof of (41), we have

$$\mathbb{P}\biggl\{ \frac{\|\boldsymbol{e}\|^2_{P_n(\alpha) - \mathbb{P}_n(\underline{\alpha}_n^c)}}{d_n(\alpha) - d_n(\underline{\alpha}_n^c)} - \sigma^2 > \frac{\sigma^2}{2} \biggr\} \leq C\biggl(\frac{\sigma^2}{2}\biggr)^{-2m_3} \{d_n(\alpha) - d_n(\underline{\alpha}_n^c)\}^{-m_3}. \tag{75}$$

Combining assumption (23) and inequality (75), there exists $n_2, k_1 \in \mathbb{N}$ such that for all $n > n_2$,

$$P\left\{\max_{\alpha\in\mathcal{A}_n^c,\, d_n(\alpha)-d_n(\underline{\alpha}_n^c)>k_1} \frac{\|\boldsymbol{e}\|^2_{P_n(\alpha)-\mathbb{P}_n(\underline{\alpha}_n^c)}}{d_n(\alpha)-d_n(\underline{\alpha}_n^c)} - \sigma^2 > \frac{\sigma^2}{2}\right\} \le \frac{\varepsilon}{2} \tag{76}$$

It follows from (74) and (76) that for all $n > \max\{n_1, n_2\}$,

$$P\left\{\max_{\alpha\in\mathcal{A}_n^c,\, d_n(\alpha)-d_n(\underline{\alpha}_n^c)>k_1} \frac{\|\boldsymbol{e}\|^2_{P_n(\alpha)-\mathbb{P}_n(\underline{\alpha}_n^c)}}{d_n(\alpha)-d_n(\underline{\alpha}_n^c)} - \sigma^2 > 2\hat{\sigma}_n^2 - \sigma^2\right\} < \varepsilon, \tag{77}$$

which further implies that

$$\bigcup_{\alpha\in\mathcal{A}_n^c,\, d_n(\alpha)-d_n(\underline{\alpha}_n^c)>k_1} \left\{G_{n,2}(\alpha) < G_{n,2}(\underline{\alpha}_n^c)\right\} \tag{78}$$

holds with probability less than ε. Therefore, $d_n(\hat{\alpha}_{\mathrm{GIC}_2}) = d_n(\underline{\alpha}_n^c) + O_p(1)$.

Step 2): We prove that $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c\} \to 1$. We first prove $\mathbb{P}\{\hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$, which only requires the proof of

$$\mathbb{P}\left\{\min_{\alpha\in\mathcal{A}_n\setminus\mathcal{A}_n^c,\, p(\alpha)\le d_n(\hat{\alpha}_{\mathrm{GIC}_2})} B_{n,\lambda_n}(\alpha) > B_{n,\lambda_n}(\underline{\alpha}_n^c)\right\} \to 1. \tag{79}$$

We then prove $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c,\ \hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$. Comparing the definitions in (2) and (5), we can calculate from (36) that

$$B_{n,\lambda_n}(\alpha) = \\ \begin{cases} \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \dfrac{\lambda_n\hat{\sigma}_n^2 H_{d_n(\alpha)}}{n} - \dfrac{\|\boldsymbol{e}_n\|^2_{P_n(\alpha)}}{n} & \text{if } \alpha\in\mathcal{A}_n^c \\ \dfrac{\|\boldsymbol{e}_n\|^2}{n} + \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} + \dfrac{\lambda_n\hat{\sigma}_n^2 H_{d_n(\alpha)} - 2\sigma^2 d_n(\alpha)}{n} & \text{if } \alpha\in\mathcal{A}_n\setminus\mathcal{A}_n^c \end{cases} \tag{80}$$

where the o_p is taken uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. Under identity (28) and $\hat{\sigma}_n^2 \to_p \sigma^2$, it was proved in (Shao, 1997, eq.(6.1)) that $|\hat{\sigma}_n^2 - \sigma^2| d_n(\alpha) = o_p\{n\mathcal{L}_n(\alpha)\}$ uniformly in $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. Thus, the second part of identity (80) may be rewritten as

$$B_{n,\lambda_n}(\alpha) = \frac{\|\boldsymbol{e}_n\|^2}{n} + \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} + \frac{\{\lambda_n H_{d_n(\alpha)} - 2d_n(\alpha)\}\hat{\sigma}_n^2}{n} \tag{81}$$

if $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$. From (80) and (81), for all $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c$,

$$\begin{aligned} B_{n,\lambda_n}(\alpha) - B_{n,\lambda_n}(\underline{\alpha}_n^c) =& \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} - \frac{2\sigma^2 d_n(\alpha)}{n} + \frac{\|\boldsymbol{e}_n\|^2_{\mathbb{P}_n(\underline{\alpha}_n^c)}}{n} \\ &+ \frac{\lambda_n \hat{\sigma}_n^2 (H_{d_n(\alpha)} - H_{d_n(\underline{\alpha}_n^c)})}{n}. \end{aligned}$$

Thus, to prove (79) it suffices to prove that

$$\mathbb{P}\left\{ \min_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c,\, p(\alpha) \le d_n(\hat{\alpha}_{\mathrm{GIC}_2})} \left(\mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} - \frac{2\sigma^2 d_n(\alpha)}{n} - \frac{\lambda_n \hat{\sigma}_n^2 H_{d_n(\underline{\alpha}_n^c)}}{n} \right) \right. \\ \left. > 0 \right\} \to 1,$$

or the following stronger result

$$\mathbb{P}\left\{ \bigcap_{\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c,\, p(\alpha) \le d_n(\hat{\alpha}_{\mathrm{GIC}_2})} \left\{ n\mathcal{L}_n(\alpha)\{1 + o_p(1)\} > 2\sigma^2 d_n(\alpha) + \lambda_n \hat{\sigma}_n^2 d_n(\underline{\alpha}_n^c) \right\} \right\} \\ \to 1. \tag{82}$$

In fact, from identity $d_n(\hat{\alpha}_{\mathrm{GIC}_2}) = d_n(\underline{\alpha}_n^c) + O_p(1)$ proved in step 1), $2\sigma^2 d_n(\alpha)$ is negligible compared with $\lambda_n \hat{\sigma}_n^2 d_n(\underline{\alpha}_n^c)$. And thus (82) is further implied by conditions (16), (29), and $\hat{\sigma}_n^2 \to_p \sigma^2$.

To prove $\mathbb{P}\{\hat{\alpha}_n = \underline{\alpha}_n^c,\ \hat{\alpha}_n \in \mathcal{A}_n^c\} \to 1$, it suffices to show that given any $\varepsilon > 0$, $\mathbb{P}(d_n(\hat{\alpha}_n) > d_n(\underline{\alpha}_n^c)\ , \hat{\alpha}_n \in \mathcal{A}_n^c) < \varepsilon$ for all sufficiently large n. From step 1) there exists a positive integer k_1 such that

$$\mathbb{P}\{d_n(\hat{\alpha}_n) - d_n(\underline{\alpha}_n^c) \ge k_1\} < \frac{\varepsilon}{2}\ . \tag{83}$$

It remains to prove that $\mathbb{P}\{0 < d_n(\hat{\alpha}_n) - d_n(\underline{\alpha}_n^c) < k_1\ , \hat{\alpha}_n \in \mathcal{A}_n^c) < \varepsilon/2$ for all sufficiently large n. From (80),

$$\begin{aligned} &\mathbb{P}\left\{ \min_{\alpha \in \mathcal{A}_n^c, 0 < d_n(\alpha) - d_n(\underline{\alpha}_n^c) < k_1} B_{n,\lambda_n}(\alpha) < B_{n,\lambda_n}(\underline{\alpha}_n^c) \right\} \\ &\le \mathbb{P}\left\{ \bigcup_{\alpha \in \mathcal{A}_n^c, 0 < d_n(\alpha) - d_n(\underline{\alpha}_n^c) < k_1} \left\{ \lambda_n \hat{\sigma}_n^2 (H_{d_n(\alpha)} - H_{d_n(\underline{\alpha}_n^c)}) < \|\boldsymbol{e}_n\|^2_{P_n(\alpha) - \mathbb{P}_n(\underline{\alpha}_n^c)} \right\} \right\} \\ &\le \mathbb{P}\left\{ \frac{\lambda_n \hat{\sigma}_n^2}{d_n(\underline{\alpha}_n^c) + 1} < \max_{\alpha \in \mathcal{A}_n^c, 0 < d_n(\alpha) - d_n(\underline{\alpha}_n^c) < k_1} \|\boldsymbol{e}_n\|^2_{P_n(\alpha) - \mathbb{P}_n(\underline{\alpha}_n^c)} \right\}. \end{aligned} \tag{84}$$

Conditions (24), (40) (with m_2 replaced by m_3), and Markov's inequality imply

$$\max_{\alpha\in\mathcal{A}_n^c, 0<d_n(\alpha)-d_n(\underline{\alpha}_n^c)<k_1} \frac{\|\boldsymbol{e}_n\|^2_{P_n(\alpha)-\mathbb{P}_n(\underline{\alpha}_n^c)}}{\{d_n(\alpha)-d_n(\underline{\alpha}_n^c)\}\hat{\sigma}_n^2} = O_p(1). \tag{85}$$

Combining (22) and (85), the value in (84) is less than $\varepsilon/2$ for all sufficiently large n. It then follows that $\mathbb{P}\{0 < d_n(\hat{\alpha}_n) - d_n(\underline{\alpha}_n^c) < k_1) < \varepsilon/2$ for all sufficiently large n.

Asymptotic loss and risk efficiency under $\mathcal{A}_n^c = \emptyset$

Given Proposition 2 and the assumptions of Theorem 2, GIC_2 is asymptotically risk efficient. Furthermore, the regularity of $\mathcal{R}_n(\cdot)$ implies that (see Durrett, 2010, Theorem 2.3.2)

$$\max_{\alpha\in\mathcal{A}_n^R}\left|\frac{d_n(\hat{\alpha}_{\text{GIC}_2})}{d_n(\alpha)} - 1\right| \to_p 0. \tag{86}$$

From (81), for all $\alpha \in \mathcal{A}_n \setminus \mathcal{A}_n^c = \mathcal{A}_n$,

$$\begin{aligned}
&B_{n,\lambda_n}(\alpha) - B_{n,\lambda_n}(\hat{\alpha}_{\text{GIC}_2}) \\
&\quad = \mathcal{L}_n(\alpha)\{1+o_p(1)\} - \mathcal{L}_n(\hat{\alpha}_{\text{GIC}_2})\{1+o_p(1)\} + \frac{\hat{\sigma}_n^2}{n}J_n(\alpha),
\end{aligned} \tag{87}$$

where $J_n(\alpha) \triangleq \lambda_n\{H_{d_n(\alpha)} - H_{d_n(\hat{\alpha}_{\text{GIC}_2})}\} - 2\{d_n(\alpha) - d_n(\hat{\alpha}_{\text{GIC}_2})\}$. Next, we prove that

$$\mathbb{P}\left\{\min_{\alpha\in\mathcal{A}_n\setminus\mathcal{A}_n^c, d_n(\alpha)<d_n(\hat{\alpha}_{\text{GIC}_2})} J_n(\alpha) > 0\right\} \to 1. \tag{88}$$

In fact, it can be verified that the function

$$g(d) \triangleq -\lambda_n \sum_{j=d+1}^{d_n(\hat{\alpha}_{\text{GIC}_2})} \frac{1}{j} - 2\{d - d_n(\hat{\alpha}_{\text{GIC}_2})\} \tag{89}$$

for a given $d_n(\hat{\alpha}_{\text{GIC}_2})$ achieves its minimum over $d = 1, \ldots, d_n(\hat{\alpha}_{\text{GIC}_2})$ at either $d = 1$ or $d = d_n(\hat{\alpha}_{\text{GIC}_2}) - 1$. Using conditions (25), (86), and the inequality $H_j < \log j + 1$ ($\forall j \in \mathbb{N}$), we can easily calculate that

$$g(1) > -\lambda_n \log d_n(\hat{\alpha}_{\mathrm{GIC}_2}) + 2\{d_n(\hat{\alpha}_{\mathrm{GIC}_2}) - 1\}$$
$$= 2d_n(\hat{\alpha}_{\mathrm{GIC}_2})\{1 + o_p(1)\}\left[\frac{d_n(\alpha_n^R) - 1}{d_n(\alpha_n^R)}\{1 + o_p(1)\} - \frac{q}{2}\right] > 0, \quad (90)$$
$$g(d_n(\hat{\alpha}_{\mathrm{GIC}_2}) - 1) = -\frac{\lambda_n}{d_n(\hat{\alpha}_{\mathrm{GIC}_2})} + 2 > 0, \quad (91)$$

with probability approaching one as $n \to \infty$. We thus proved (88) from (90) and (91).

From the definition of of $\hat{\alpha}_n$, $B_{n,\lambda_n}(\hat{\alpha}_n) - B_{n,\lambda_n}(\hat{\alpha}_{\mathrm{GIC}_2}) \leq 0$. It then follows from (87) and (88) that

$$\mathbb{P}\left\{\mathcal{L}_n(\hat{\alpha}_n)\{1 + o_p(1)\} - \mathcal{L}_n(\hat{\alpha}_{\mathrm{GIC}_2})\{1 + o_p(1)\} \leq 0\right\} \to 1. \quad (92)$$

Dividing both sides of (92) by $\mathcal{L}_n(\hat{\alpha}_{\mathrm{GIC}_2})$, we obtain that for any fixed $\varepsilon > 0$,

$$\mathbb{P}\left\{\frac{\mathcal{L}_n(\hat{\alpha}_n)}{\mathcal{L}_n(\hat{\alpha}_{\mathrm{GIC}_2})} < 1 + \varepsilon\right\} \to 1. \quad (93)$$

Combining (93) with $\mathcal{L}_n(\hat{\alpha}_{\mathrm{GIC}_2})/\mathcal{L}_n(\alpha_n^L) \to_p 1$, we obtain for any fixed $\varepsilon > 0$ that

$$\mathbb{P}\left\{\frac{\mathcal{L}_n(\hat{\alpha}_n)}{\mathcal{L}_n(\alpha_n^L)} < 1 + \varepsilon\right\} \to 1. \quad (94)$$

On the other hand, by the definition of α_n^L, $\mathcal{L}_n(\hat{\alpha}_n) \geq \mathcal{L}_n(\alpha_n^L)$. Thus, we have proved the asymptotic loss efficiency of $\hat{\alpha}_n$. The asymptotic risk efficiency directly follows from (29).

H. Proof of Proposition 3

In the parametric scenario, the consistency of BC and GIC_{λ_n} indicates that $\mathrm{PI}_n \to_p 1$. In the nonparametric scenario, we obtain from the regularity of $\mathcal{R}_n(\cdot)$, the definition of α_n^R, and the efficiency of $\hat{\alpha}_{\mathrm{GIC}_2}, \hat{\alpha}_{\mathrm{BC}}$ that

$d_n(\hat{\alpha}_{\text{GIC}_2})/d_n(\alpha_n^R), d_n(\hat{\alpha}_{\text{BC}})/d_n(\alpha_n^R) \to_p 1$, which further implies

$$d_n(\hat{\alpha}_{\text{BC}})/d_n(\hat{\alpha}_{\text{GIC}_2}) \to_p 1.$$

We shall also prove that

$$\mathbb{P}\left\{\frac{d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}})}{d_n(\hat{\alpha}_{\text{GIC}_2})} < c_2\right\} \to 1 \tag{95}$$

as $n \to \infty$. Then, with probability tending to one, we have

$$\text{PI}_n \leq \frac{|d_n(\hat{\alpha}_{\text{BC}}) - d_n(\hat{\alpha}_{\text{GIC}_2})|}{|d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}}) - d_n(\hat{\alpha}_{\text{GIC}_2})|} = \frac{|d_n(\hat{\alpha}_{\text{BC}})/d_n(\hat{\alpha}_{\text{GIC}_2}) - 1|}{|d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}})/d_n(\hat{\alpha}_{\text{GIC}_2}) - 1)|} \to_p 0,$$

which concludes the proof.

To prove (95), we rewrite $G_{n,\lambda_n}(\alpha)$ (for $\alpha \in \mathcal{A}_n$) in (36) as

$$\begin{aligned} G_{n,\lambda_n}(\alpha) =& \frac{\|\boldsymbol{e}_n\|^2}{n} + \mathcal{L}_n(\alpha) + o_p\{\mathcal{L}_n(\alpha)\} + \frac{(\lambda_n - 2)\sigma^2 d_n(\alpha)}{n} \\ &+ \frac{\lambda_n(\hat{\sigma}_n^2 - \sigma^2)d_n(\alpha)}{n}. \end{aligned} \tag{96}$$

By the assumption $\hat{\sigma}_n^2 \to_p \sigma^2$ and the first inequality in (26), we have $\lambda_n(\hat{\sigma}_n^2 - \sigma^2)d_n(\alpha)/n = o_p(\mathcal{R}_n(\alpha))$ uniformly in $\alpha \in \mathcal{A}_n$. From (29), we may rewrite (96) as

$$\begin{aligned} G_{n,\lambda_n}(\alpha) &= \frac{\|\boldsymbol{e}_n\|^2}{n} + \{1 + o_p(1)\}\mathcal{R}_n(\alpha) + (\lambda_n - 2)\sigma^2 d_n(\alpha)/n \\ &= C + \{1 + o_p(1)\}\mathcal{R}_n^*(\alpha) \end{aligned}$$

where $C = \|\boldsymbol{e}_n\|^2/n$ does not depend on α. By the definition of $\hat{\alpha}_{\text{GIC}_{\lambda_n}}, \alpha_n^*$, it follows that

$$\frac{\mathcal{R}_n^*(\hat{\alpha}_{\text{GIC}_{\lambda_n}})}{\mathcal{R}_n^*(\alpha_n^*)} \to_p 1,$$

and further from the regularity of $\mathcal{R}_n^*(\cdot)$, $d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}})/d_n(\alpha_n^*) \to_p 1$. Therefore, we obtain

$$\frac{d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}})}{d_n(\hat{\alpha}_{\text{GIC}_2})} = \frac{d_n(\hat{\alpha}_{\text{GIC}_{\lambda_n}})}{d_n(\alpha_n^*)} \frac{d_n(\alpha_n^*)}{d_n(\alpha_n^R)} \frac{d_n(\alpha_n^R)}{d_n(\hat{\alpha}_{\text{GIC}_2})} = \{1 + o_p(1)\}\frac{d_n(\alpha_n^*)}{d_n(\alpha_n^R)} < c_2$$

with probability tending to one as $n \to \infty$.

I. Another definition of PI

Another definition of PI was given by Liu and Yang (2011). It is defined by

$$\text{PI}_n^{(2)} = \min_{\alpha \in \Lambda(\hat{\alpha}_n)} \frac{S_n(\alpha) + d_n(\alpha)\hat{\sigma}_n^2 \log n - n\hat{\sigma}_n^2}{S_n(\hat{\alpha}_n) + d_n(\hat{\alpha}_n)\hat{\sigma}_n^2 \log n - n\hat{\sigma}_n^2}$$

if $d_n(\hat{\alpha}_n) > 1$, and $\text{PI}_n^{(2)} = n$ otherwise, where $\hat{\alpha}_n$ is selected by some consistent procedure (in the sense that it is consistent in the parametric scenario), and $\Lambda(\alpha)$ is the set of sub-models of α with dimension $d_n(\alpha) - 1$. It is assumed in (Liu and Yang, 2011) that $\Lambda(\alpha)$ is nonempty for any α with dimension greater than one. In our simulation, we relax the definition to be the set of sub-models whose dimension is the closest (but not equal) to $d_n(\alpha)$.

The intuition is that a data-generating model should be very different (in terms of goodness of fit) from any sub-model, while for a mis-specified model class, the selected model varies slightly when few variables are dropped. Liu and Yang (2011) showed that $\text{PI}_n^{(2)}$ converges in probability to ∞ and 1 in parametric and nonparametric scenarios, respectively. Our PI_n is defined based on fundamentally different motivations, and validated under different assumptions. In Section B, we shall show that both PI_n and $\text{PI}_n^{(2)}$ work as expected in various experiments.

J. Proof of Theorem 3

We consider the (typical) case where $\hat{\alpha}_{\text{BC}}$ is not equal to $\bar{\alpha}_n = \{1, \ldots, d_n\}$. Similar proof can be applied to the case where $\hat{\alpha}_{\text{BC}} = \bar{\alpha}_n$ and $\mathcal{A}_n$ is expanded during the simulated validation. For notational convenience, "stage 1" refers to the estimation and selection procedure on the original data (i.e. with mean function $\boldsymbol{f}_n$), and "stage 2" refers to the counterpart during simulated validation (i.e. with mean function $\tilde{\boldsymbol{f}}_n$).

We first prove for the parametric scenario. By Theorem 2, $\mathbb{P}(\hat{\alpha}_{\text{BC}} = \underline{\alpha}_n^c) \to 1$. We only need to verify that with $\hat{\alpha}_{\text{BC}} = \underline{\alpha}_n^c$, the conditions in Case 1 of Theorem 2 still hold at stage 2 with probability going to one. It is easy to verify that conditions (8), (22)-(24), and the estimated variance $\hat{\sigma}_s^2 \to_p \sigma^2$ hold at stage 2. Also, condition (9) would be implied by the combination of (16) and (22). Thus, it remains to prove that (13) and (16) hold with probability going to one at stage 2.

Let $\tilde{\mathcal{R}}_n(\cdot)$ denote the risk function at stage 2. Recall from (12) that $\mathcal{R}(\alpha)$ and $\tilde{\mathcal{R}}_n(\alpha)$ may be written as

$$
\begin{aligned}
n\mathcal{R}(\alpha) &= n\Delta_n(\alpha) + \sigma^2 d_n(\alpha), \\
n\tilde{\mathcal{R}}(\alpha) &= n\tilde{\Delta}_n(\alpha) + \sigma^2 d_n(\alpha).
\end{aligned}
$$

By a similar argument used to derive (42),

$$
\|\boldsymbol{f}_n - \tilde{\boldsymbol{f}}_n\|^2 = \|P_n(\underline{\alpha}_n^c)\tilde{\boldsymbol{\varepsilon}}_n\|^2 = O_p(d_n(\underline{\alpha}_n^c)).
$$

The difference between $\tilde{\Delta}_n(\alpha)$ and $\Delta_n(\alpha)$ can be uniformly bounded by

$$
\begin{aligned}
\max_{\alpha \in \mathcal{A}_n} \left| \sqrt{n\tilde{\Delta}_n(\alpha)} - \sqrt{n\Delta_n(\alpha)} \right| &= \max_{\alpha \in \mathcal{A}_n} \Big| \|P_n(\alpha)^{\perp} \tilde{\boldsymbol{f}}_n\| - \|P_n(\alpha)^{\perp} \boldsymbol{f}_n\| \Big| \\
&\le \max_{\alpha \in \mathcal{A}_n} \|P_n(\alpha)^{\perp}(\tilde{\boldsymbol{f}}_n - \boldsymbol{f}_n)\| \\
&\le \|\tilde{\boldsymbol{f}}_n - \boldsymbol{f}_n\| \\
&= O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big). \qquad (97)
\end{aligned}
$$

Therefore, we have

$$\begin{aligned}
\max_{\alpha\in\mathcal{A}_n\setminus\mathcal{A}_n^c}|\tilde{\mathcal{R}}(\alpha)-\mathcal{R}(\alpha)| &= \max_{\alpha\in\mathcal{A}_n\setminus\mathcal{A}_n^c}|n\tilde{\Delta}_n(\alpha)-n\Delta_n(\alpha)|\\
&\le O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big)\max_{\alpha\in\mathcal{A}_n}\bigg(\sqrt{n\tilde{\Delta}_n(\alpha)}+\sqrt{n\Delta_n(\alpha)}\bigg)\\
&\le O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big)\bigg(O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big)+2\sqrt{n\Delta_n(\alpha)}\bigg)\\
&< O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big)\bigg(O_p\big(\sqrt{d_n(\underline{\alpha}_n^c)}\big)+2\sqrt{\mathcal{R}(\alpha)}\bigg)\\
&= o_p(1)\mathcal{R}(\alpha) \qquad (98)
\end{aligned}$$

with probability going to one as $n\to\infty$, where (98) is due to condition (9) at stage 1. It follows from (98) that conditions (13) and (16) hold with probability going to one at stage 2.

We then prove for the nonparametric scenario. Suppose that $\hat{\alpha}_{\text{BC}}$ is nested under a larger model α' with dimension $d_n(\hat{\alpha}_{\text{BC}})+1$. Let $\hat{\alpha}_{\text{GIC}_2,s}$ denote the model selected by AIC in stage 2. We rewrite $\mathbb{P}(\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}})$ as

$$\begin{aligned}
\mathbb{P}(\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}}) &= p_1+p_2+p_3, \qquad (99)\\
p_1 &= \mathbb{P}(\hat{\alpha}_{\text{GIC}_2,s}>\hat{\alpha}_{\text{BC}},\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}}),\\
p_2 &= \mathbb{P}(\hat{\alpha}_{\text{GIC}_2,s}=\hat{\alpha}_{\text{BC}},\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}}),\\
p_3 &= \mathbb{P}(\hat{\alpha}_{\text{GIC}_2,s}<\hat{\alpha}_{\text{BC}},\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}}).
\end{aligned}$$

By the rule of BC in (5), $p_3=0$. Next we bound p_1 and p_2.

The event $\hat{\alpha}_{\text{GIC}_2,s}>\hat{\alpha}_{\text{BC}},\hat{\alpha}_{\text{BC},s}=\hat{\alpha}_{\text{BC}}$ implies that BC favors $\hat{\alpha}_{\text{BC}}$ over α', namely

$$\tilde{S}_n(\hat{\alpha}_{\text{BC}})-\tilde{S}_n(\alpha')<\frac{\lambda_n\hat{\sigma}_s^2}{d_n(\hat{\alpha}_{\text{BC}})+1}$$

by the definition in (5). Here, $\tilde{S}(\alpha)\overset{\Delta}{=}\|\tilde{\boldsymbol{y}}_n-\hat{\boldsymbol{f}}_{n,s}(\alpha)\|^2$, and $\hat{\boldsymbol{f}}_{n,s}$ is the least square estimates of $\tilde{\boldsymbol{f}}_n$. As a result,

$$p_1\le\mathbb{P}\bigg(\hat{\alpha}_{\text{GIC}_2,s}>\hat{\alpha}_{\text{BC}},E_1\bigg)$$

where E_1 denotes the event

$$\frac{\tilde{S}_n(\hat{\alpha}_{\mathrm{BC},s}) - \tilde{S}_n(\alpha')}{\sigma^2} < \frac{\lambda_n \hat{\sigma}_s^2}{\sigma^2\{d_n(\hat{\alpha}_{\mathrm{BC}}) + 1\}}. \tag{100}$$

The left hand term in (100) may be rewritten as

$$C \triangleq \sigma^{-2} \|\boldsymbol{e}_s\|^2_{P_n(\alpha') - P_n(\hat{\alpha}_{\mathrm{BC}})}, \tag{101}$$

which follows chi square distribution with one degree of freedom (χ_1^2) since e_s is Gaussian in the simulated validation. The second hand term in (100) is

$$\frac{\hat{\sigma}_s^2}{\sigma^2} \frac{\lambda_n}{d_n(\alpha_n^R)} \frac{d_n(\alpha_n^R)}{d_n(\hat{\alpha}_{\mathrm{BC}}) + 1}$$

which, by applying $\hat{\sigma}_s^2 \to_p \sigma^2$, condition (25) and $d_n(\hat{\alpha}_{\mathrm{BC}})/d_n(\alpha_n^R) \to_p 1$, is asymptotically less than 2. Therefore, we have

$$p_1 \leq \mathbb{P}(\hat{\alpha}_{\mathrm{GIC}_2,s} > \hat{\alpha}_{\mathrm{BC}}, C < 3) \tag{102}$$

Likewise, the event $\hat{\alpha}_{\mathrm{GIC}_2,s} = \hat{\alpha}_{\mathrm{BC}}, \hat{\alpha}_{\mathrm{BC},s} = \hat{\alpha}_{\mathrm{BC}}$ implies that

$$\tilde{S}_n(\hat{\alpha}_{\mathrm{BC}}) - \tilde{S}_n(\alpha') < 2\hat{\sigma}_s^2$$

by the definition in (2). This implies that

$$p_2 \leq \mathbb{P}(\hat{\alpha}_{\mathrm{GIC}_2,s} = \hat{\alpha}_{\mathrm{BC}}) \leq \mathbb{P}(\hat{\alpha}_{\mathrm{GIC}_2,s} = \hat{\alpha}_{\mathrm{BC}}, C < \frac{2\hat{\sigma}_s^2}{\sigma^2} < 3) \tag{103}$$

with probability close to one for large n. Combining (99), (102) and (103), we obtain $\mathbb{P}(\hat{\alpha}_{\mathrm{BC},s} = \hat{\alpha}_{\mathrm{BC}}) \leq \mathbb{P}(C < 3)$ where C is a χ_1^2 random variable. This concludes the proof.

K. Theoretical analysis of BC-VO

We introduce some helpful notation for the theoretical results. Let X_1, X_2 denote the design matrices corresponding to the two disjoint datasets $\mathcal{D}_1, \mathcal{D}_2$, respectively. Let $X_1(\underline{\alpha}_n^c), X_1(-\underline{\alpha}_n^c)$ denote the submatrices of X_1 consisting of columns in $\underline{\alpha}_n^c$, $\{1, \ldots, d_n\} - \underline{\alpha}_n^c$, respectively. Let $X_1(k)$ denote the kth column of X_1. For a matrix $M \in \mathbb{R}^{u \times v}$ Let $\|M\|_\infty = \max_{1 \leq i \leq u} \sum_{1 \leq j \leq v} |M_{ij}|$. We standardize the data in $\mathcal{D}_{1a}$ such that

$$n^{-1/2} \max_{j \in \{1,\ldots,d_n\} \setminus \underline{\alpha}_n^c} \|X_1(j)\| \leq 1.$$

Similar standardization is applied to $\mathcal{D}_{\text{te}}$.

K.1. ARM weighting scheme

As a default option in 'bc' package, we use the following ARM weighting procedure.

- Randomly split $\mathcal{D}_{1b}$ into a training set $\mathcal{D}_{\text{tr}}$ and a testing set $\mathcal{D}_{\text{te}}$;
- For each model $\alpha \in \hat{\mathcal{A}}_n$, we use the least squares method on $\mathcal{D}_{\text{tr}}$ to obtain an estimated coefficients $\hat{\boldsymbol{\beta}}_\alpha$ (and the corresponding regression function $\hat{f}_\alpha(\cdot)$) and noise variance $\hat{\sigma}^2$;
- For each model $\alpha \in \hat{\mathcal{A}}_n$, compute the (unnormalized) weight

$$w_\alpha = e^{-C_\alpha} \prod_{i \in \mathcal{D}_{\text{te}}} N(y_i; \hat{f}_\alpha(\boldsymbol{x}_i), \hat{\sigma}^2), \text{ where} \tag{104}$$

$$C_\alpha = d_n(\alpha) \log\bigl(1 + \frac{d_n}{d_n(\alpha)}\bigr) + 2\log(d_n(\alpha) + 2)$$

  and $N(y; \mu, \sigma^2)$ denotes the density of Gaussian distribution with mean μ and variance σ^2 evaluated at y.

Note that $\mathcal{D}_{1b}$ is used for parameter estimation by each model and $\mathcal{D}_{1a}$ is used to assess the prediction performance and then the weights are assigned accordingly. Compared with the original ARM by Yang (2001), the model-specific variance $\hat{\sigma}_\alpha^2$ in (104) is replaced with $\hat{\sigma}^2$ that can be estimated using all the variables in $\hat{\mathcal{A}}_n$. This is mainly for technical convenience, and has little effect on empirical performance from our experimental studies.

K.2. Accuracy of variable ordering

Proposition 5 below implies that the first s variables in variable ordering exactly match all the significant variables. It further implies that $\underline{\alpha}_n^c \in \mathcal{A}_n$.

PROPOSITION 5. *For the variable importance u_k's produced by step (d), $\underline{\alpha}^c \in \hat{\mathcal{A}}_n$ implies that*

$$\min_{k\in\underline{\alpha}^c} u_k > \max_{j\notin\underline{\alpha}^c} u_j$$

with probability going to one as $n \rightarrow \infty$, given that the following conditions hold.

1. In step (b), the number of subsets from each penalized method is restricted to be no more than $c_0/3$, and $\max_{\alpha\in\hat{\mathcal{A}}_n} d(\alpha) \leq c_0$ for some constant $c_0 > 3s$ that does not depend on n.

2. In step (c), $\mathcal{D}_{1b}\cap\mathcal{D}_{1a} = \emptyset$, and $card(\mathcal{D}_{te})/card(\mathcal{D}_{tr}) \rightarrow 0$, $card(\mathcal{D}_{te}) \rightarrow \infty$ as $n \rightarrow \infty$ in the ARM weighting scheme.

3. There exists a constant δ_{c_0} such that $\|X_{tr}\boldsymbol{\beta}\| \geq \delta_{c_0}\|\boldsymbol{\beta}\|^2$ for every $\boldsymbol{\beta} \in \mathbb{R}^p$ with at most c_0 nonzeros.

Condition 1 is only for technical convenience. For LASSO, it has been proved under some conditions that the solution paths are piecewise linear and nested (Efron et al., 2004; Rosset and Zhu, 2007; Tibshirani et al., 2013). It is easy to control the number of selected variables (and thus the number of candidates) through the regularization parameter.

For SCAD and MCP, it seems not clear whether the produced solution paths are nested up to the authors' knowledge. In practice, we found through numerical experiments that the above lemma generally holds with $c_0 = 3\sqrt{n_{1a}}$ for a broad class of Gaussian ensembles. And this is the default parameter in 'bc' package. Condition 2 is specific to the weighting scheme. The intuition is that training size needs to be sufficiently large so that the weight of the smallest correct model $\underline{\alpha}_n^c$ is not negligible compared with larger models that nest it. Condition 3 is to guarantee that the least squares method is applicable and the model $\underline{\alpha}^c$ has a non-negligible weight. This condition can be implied by, e.g., the existence of c_0-restricted isometry property (Candes and Tao, 2005).

Proof of Proposition 5:

For brevity, let $n_{\mathrm{tr}} = \mathrm{card}(\mathcal{D}_{\mathrm{tr}})$, $n_{\mathrm{te}} = \mathrm{card}(\mathcal{D}_{\mathrm{te}})$. For any $\alpha \in \hat{\mathcal{A}}_n \cap \mathcal{A}_n^c$, Condition 3 implies that $\underline{\alpha}^c \subseteq \alpha$. Therefore, by definition of u_k's in step (d) it remains to prove

$$\max_{\alpha \in \hat{\mathcal{A}}_n \setminus \mathcal{A}_n^c} \frac{w_\alpha}{w_{\underline{\alpha}^c}} \to_p 0, \qquad \max_{\alpha \in \hat{\mathcal{A}}_n \cap \mathcal{A}_n^c} \frac{w_\alpha}{w_{\underline{\alpha}^c}} < c, \tag{105}$$

for some fixed constant c with probability going to one, as $n \to \infty$. The w_α in (104) may be rewritten as

$$w_\alpha \propto c_\alpha \exp\biggl\{-\frac{1}{2\hat{\sigma}^2}\|X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*) - \boldsymbol{e}\|^2\biggr\}$$

where $\boldsymbol{e}$ is the i.i.d. noise vector of size n_{te}, and $\hat{\boldsymbol{\beta}}_\alpha$ is the least squares estimate from $\mathcal{D}_{\mathrm{tr}}$.

Thus, to prove (105) it is sufficient to prove that for any $\alpha \in \hat{\mathcal{A}}_n - \mathcal{A}_n^c$

$$B \triangleq \|X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_{\underline{\alpha}^c} - \boldsymbol{\beta}_*) - \boldsymbol{e}\|^2 - \|X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*) - \boldsymbol{e}\|^2 \tag{106}$$

goes to $-\infty$ in probability, and for any $\alpha \in \hat{\mathcal{A}}_n \cap \mathcal{A}_n^c$, B converges to zero in probability, as $n \to \infty$. We rewrite (106) as $B_1 + B_2 + +B_3$ where

$$B_1 = \|X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_{\underline{\alpha}^c} - \boldsymbol{\beta}_*)\|^2, \quad B_2 = \|X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*)\|^2, \quad B_3 = 2\boldsymbol{e}^{\mathrm{T}}X_{\mathrm{te}}(\hat{\boldsymbol{\beta}}_{\underline{\alpha}^c} - \hat{\boldsymbol{\beta}}_\alpha).$$

It follows from Condition 2 and $B_1 = (n_{\mathrm{te}}/n_{\mathrm{tr}})\|n_{\mathrm{te}}^{-1/2}X_{\mathrm{te}}\ \delta_\alpha\|^2$ where $\delta_\alpha \overset{\Delta}{=} \sqrt{n_{\mathrm{tr}}}(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_*) = O_p(1)$ that $B_1 = o_p(1)$. For $\alpha \in \hat{\mathcal{A}}_n - \mathcal{A}_n^c$, Condition 3 implies that $B_2 = O_p(n_{\mathrm{te}})$ and $B_2 = O_p(n_{\mathrm{te}}^{1/2})$. Thus $B \to_p -\infty$ as $n \to \infty$. For $\alpha \in \hat{\mathcal{A}}_n \cap \mathcal{A}_n^c$, it can be similarly proved that $B_2 = o_p(1)$ and $B_3 = o_p(1)$, and thus $B \to_p 0$ as $n \to \infty$.

K.3. Oracle property of BC-VO

The following Theorem 4 establish the oracle property of BC in sparse subset selection. Note that we assume fixed design matrix for technical convenience. Extensions to the case of random design matrices are possible, but we will not elaborate in the rest of the paper.

THEOREM 4. *The above step (f) produces an estimate that satisfies the oracle property and $\mathcal{R}$-consistency, assuming that the conditions of Proposition 5 hold and the following conditions hold.*

1. *There exists some fixed constant $c_1 \in (0, 1]$ such that*

$$\left\| X_{1a}(-\underline{\alpha}^c)^{\mathrm{T}} X_{1a}(\underline{\alpha}^c) \left\{ X_{1a}(\underline{\alpha}^c)^T X_{1a}(\underline{\alpha}^c) \right\}^{-1} \right\|_\infty \leq 1 - c_1.$$

2. *There exists some fixed constant $c_2 > 0$ such that*

$$eig_{\min}\left(\frac{1}{n} X_{1a}(\underline{\alpha}^c)^T X_{1a}(\underline{\alpha}^c)\right) \geq c_2.$$

3. *There exists a constant $\delta^{'}_{c_0}$ such that $\|X_2\boldsymbol{\beta}\| \geq \delta^{'}_{c_0}\|\boldsymbol{\beta}\|^2$ for every $\boldsymbol{\beta} \in \mathbb{R}^p$ with at most c_0 nonzeros.*

4. *$\lambda_n \to \infty$.*

5. *$E(e_1^4) < \infty$.*

Conditions 1 and 2 are typical assumptions for high dimensional regression, and they were also considered in e.g. Fuchs (2005); Wainwright (2009). Condition 3 here is similar to Condition 3 of Proposition 5, and they are usually guaranteed with high probability for randomly generated design matrices (Candes and Tao, 2005). Its use here is to distinguish the "parametric" from the "nonparametric" models. Condition 4 is a mild requirement of choosing λ_n. No upper bound of λ_n is required because we are only interested in selection consistency here. Condition 5 is a mild regularity assumption on noises.

Proof of Theorem 4:

Given Conditions 1 and 2, and β_* being fixed, it follows from (Wainwright, 2009, Theorem 1) that $\underline{\alpha}^c$ can be selected by LASSO with probability going to one for some regularization parameter. Given $c_0/3 > s$, this further implies that $\mathbb{P}(\underline{\alpha}_n^c \in \hat{\mathcal{A}}_n) \to 1$ as $n \to \infty$. Therefore, by using Proposition 5 we obtain

$$\mathbb{P}(\underline{\alpha}^c \in \mathcal{A}_n) \to 1 \tag{107}$$

as $n \to \infty$ in step (e). It remains to show that the conditions of Case 1 in Theorem 2 hold when applying step (f).

Because $\mathcal{A}_n$ is nested and finite, (107) implies that $\hat{\sigma}_n^2 \to_p \sigma^2$ and condition (8). Condition 3 implies (9) and (13), and Condition 4 implies (22). Moreover, (23) and (24) are trivially satisfied with $m_1 = 1$. By applying Theorem 2 we obtain the consistency in selecting $\underline{\alpha}^c$.

In step (f), with $\underline{\alpha}^c$ being correctly selected, one could estimate β_* at oracle rates by restricting to $\underline{\alpha}^c$. To prove $\mathcal{R}$-consistency, we observe from (12) that any model α with $\mathcal{R}_n(\alpha) \leq \mathcal{R}_n(\underline{\alpha}^c) = \sigma^2 s/n$ must satisfy $d_n(\alpha) \leq s$. By Condition 3 of Theorem 4 and $c_0 \geq 3s$, we have $\Delta_n(\alpha) > \delta$ for some fixed constant $\delta > 0$. This is impossible as $n \to \infty$.