Statistica Sinica **34** (2024), 505-521
doi:https://doi.org/10.5705/ss.202021.0281

MEASURES OF UNCERTAINTY FOR SHRINKAGE MODEL SELECTION

Yuanyuan Li and Jiming Jiang*

University of California

Abstract: We develop measures of uncertainty, including model confidence sets and a LogP measure, for shrinkage model selection procedures. The measures are developed for linear models, generalized linear models, and generalized additive models. We study the theoretical and empirical properties of the proposed measures, and demonstrate how theses measures work by applying them to real-life problems.

Key words and phrases: Asymptotic coverage probability, average probability of coverage, consistency, nested MCS, shrinkage model selection, uncertainty.

1. Introduction

Driven largely by practical needs, measures of uncertainty in model selection have been studied by, among others, Hansen, Lunde and Nason (2011), Ferrari and Yang (2015), Lubke and Campbell (2016) (also, see Lubke et al. (2017) and the references therein), Zheng, Ferrari and Yang (2019), Li et al. (2019), and Liu, Li and Jiang (2021). For example, Conlon et al. (2003) discussed a motif regression problem. One of their study objectives is to find binding sites in DNA sequences of an NDD1 transcriptional activator, which is essential for the expression of a set of late-S-phase-specific genes. Pang, Lin and Jiang (2016) formulated these study objectives as a shrinkage variable selection problem, and several models were proposed based on various selection criteria. We revisit this problem later. As another example, Subramanian et al. (2005) proposed using a gene set enrichment analysis (GSEA) to assess the significance of predefined gene sets. Efron and Tibshirani (2007) proposed an alternative method, called a gene-set analysis (GSA), that produced different selection results to those of the GSEA when applied to the p53 data with the catalog of 522 gene sets in Subramanian et al. (2005). Quite often in practice, researchers use different variables and models when applying model-selection procedures, and it is not clear which is most suitable when the selection results vary with different procedures or methods. For example, standard model selection procedures include information criteria (e.g., AIC (Akaike (1973)), BIC (Schwarz (1978))), the fence methods (e.g., Jiang and Nguyen (2016)), and shrinkage selection/estimation (e.g., the least absolute shrinkage and selection operator (lasso) (Tibshirani (1996)), and

*Corresponding author.

smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)) estimators).

Alternatively, instead of focusing on a single model that may correspond to a subset of selected variables, one may consider a few models as possibilities. These models are all appropriate because they are the results of model selection based on different considerations. There may be other ways to justify these models. Such a group of models naturally form a model set, which may be associated with a model confidence set (MCS). Several MCSs have been proposed; see the references mentioned in the first paragraph of this section. However, note that although an MCS is typically defined as having a designed probability of covering an optimal model (see below), in practice, the latter may not exist among the candidate models under consideration. Nevertheless, an MCS may still be useful, especially when there is uncertainty, about a single selected model.

MCS offers one type of measure of uncertainty in model selection. Another type measures the error in model selection. Liu, Li and Jiang (2021) proposed a LogP measure that is an estimated logarithm of the probability of selecting a nonoptimal model. Here, an optimal model, denoted as M_{opt}, is defined as the most parsimonious correct model, typically measured in terms of having the minimum number of model parameters. Similarly to an MCS, when a correct model may not exist, the optimal model may be understood in a broader sense as one that best approximates the nature that generates the data. This potentially extends the usefulness of LogP.

Thus far, MCSs and LogP have been developed under the framework of classical model selection, where we have a finite set of candidate models with cardinality that does not increase with the sample size. Here, the consistency of the model selection is typically established for the selected model according to a certain model selection criterion. For example, Liu, Li and Jiang (2021) studied both the finite-sample and the asymptotic behavior of their proposed MCS and LogP measures. However, such methods may not apply to modern model selection problems, which are characterized by high dimensionality. It is known that shrinkage model selection (Tibshirani (1996)) methods are often suitable for high-dimensional variable selection problems. The MCS method called Model confidence bounds (MCB) of Li et al. (2019) incorporates shrinkage methods, making it promising for high-dimensional model selection. However, no studies have examined the performance of MCB in high-dimensional cases and for models beyond a linear regression, whether empirically or theoretically. Thus, the main purpose of this work is to develop MCS and LogP strategies that are suitable for shrinkage model selection, including linear models, generalized linear models (GLMs), and generalized additive models (GAMs). We also compare the performance of the proposed MCS method with that of MCB in high-dimensional cases using simulation studies.

In Section 2, we develop the MCS and LogP measures for linear models, GLMs, GAMs, and, in Section 3, we examine the properties of the proposed measures. In Section 4, we investigate the finite-sample performance of the proposed measures using Monte Carlo simulation studies. Two real-data applications are discussed in Section 5. The proofs of the theoretical results are deferred to the online Supplementary Material.

2. Methods

We first develop the measures under a linear model setting, before considering extensions to other general models.

2.1. Constructing a nested MCS for M_{opt}

2.1.1. Linear models

Suppose the data are generated under a linear model,

$$y = X\beta + \varepsilon, \tag{2.1}$$

where y is an $n \times 1$ vector of responses, X is an $n \times p$ known matrix of covariates, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of independent and identically distributed (i.i.d.) errors with mean 0 and variance σ^2. Consider a lasso solution fitting the model

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1, \tag{2.2}$$

where $\lambda > 0$ denotes the penalty or regularization parameter. Efron et al. (2004) proposed the least angle regression (LARS) algorithm to compute the full solution path of the lasso. As λ decreases from a large value, the covariates enter the model at a certain order, say EO $= \{e_1, e_2, \ldots, e_p\}$, where EO means "Entering Order". For any fixed λ, we have a corresponding selected model, $\hat{M} = \{e_1, \ldots, e_k\}$, that is, we select the first k covariates in EO. Different λ will result in different numbers of covariates in the selected model. However, the order of the covariates entering the model is fixed, and depends only on the data. Usually, λ is chosen using cross-validation or an information criterion; see, for example, Fu, Carroll and Wang (2005) and Wang, Li and Leng (2009). In rare cases, some covariates enter the model and leave multiple times, usually because of the strong correlation between covariates, and we define EO using a two-step approach based on the selected model $\hat{M}$. A detailed description and discussion are provided in the Supplementary Material (see Section S1).

Let $\hat{M}$ denote the model selected using the original data, and $\hat{\psi}$ be the vector of estimated parameters under $\hat{M}$. Later, we investigate the model using bootstrapped data. Furthermore, note that $\hat{M}$ is the first q elements in EO, for some positive integer q. Because EO may be viewed as an order of relative importance among the covariates, we may eliminate relatively unimportant

members from $\hat{M}$ to form a "smaller" model, $\hat{M}_{\mathrm{L}}$, called a lower bound model (LBM). Note that this is different to the LBM of Ferrari and Yang (2015), even though the same abbreviation is used. Similarly, adding more covariates in EO to $\hat{M}$ results in a "larger" model, $\hat{M}_{\mathrm{U}}$, called an upper bound model (UBM). It is easy to see that $\hat{M}_{\mathrm{L}} \subseteq \hat{M} \subseteq \hat{M}_{\mathrm{U}}$. If the pair of models $\{\hat{M}_L, \hat{M}_U\}$ satisfies

$$\mathrm{P}(\hat{M}_L \subseteq M_{\mathrm{opt}} \subseteq \hat{M}_U) \geq 1 - \alpha, \tag{2.3}$$

where M_{opt} is the optimal model, that is, the most parsimonious true model under which data can be generated from the same distribution as that of the observed data (Liu, Li and Jiang (2021)), then $\{M : \hat{M}_L \subseteq M \subseteq \hat{M}_U, M \subset \mathcal{M}\}$ is a $100(1-\alpha)\%$ MCS for M_{opt}, denoted as $(\hat{M}_L, \hat{M}_U)$. We call this a nested MCS (NMCS), because the construction is based on a nested sequence of models; the resulting MCS is a nested sequence between the LBM and the UBM.

It remains to determine how many covariates should be eliminated, and added, to form the LBM and UBM, respectively, so that (2.3) holds. Because M_{opt} and ψ_{opt} are unknown, the probability in (2.3) cannot be calculated directly. Following Liu, Li and Jiang (2021), we can approximate the probability in (2.3) using bootstrapping. Furthermore, an NMCS should include as few models as possible in order to be efficient. As such, we define the width of an NMCS as

$$w(\hat{M}_{\mathrm{L}}, \hat{M}_{\mathrm{U}}) = |\hat{M}_{\mathrm{U}}| - |\hat{M}_{\mathrm{L}}|, \tag{2.4}$$

where $|M|$ denotes the number of covariates included in model M. For a given width w, we determine the NMCS as follows:

$$\mathrm{NMCS}(w) = \underset{(\hat{M}_1, \hat{M}_2)}{\mathrm{argmax}} \{\mathrm{P}(\hat{M}_1 \subseteq M_{\mathrm{opt}} \subseteq \hat{M}_2) : w(\hat{M}_1, \hat{M}_2) = w, \hat{M}_1 \subseteq \hat{M} \subseteq \hat{M}_2\}, \tag{2.5}$$

where the probability P is approximated using bootstrapping (see below). In addition, because the width of an NMCS is a positive integer, we only need to consider (2.4) for w that are positive integers. The coverage probability of the NMCS is then a function of w,

$$\mathrm{CP}(w) = \mathrm{P}\{M_{\mathrm{opt}} \in \mathrm{NMCS}(w)\}. \tag{2.6}$$

Suppose that the optimal model satisfies $M_{\mathrm{null}} \subseteq M_{\mathrm{opt}} \subseteq M_{\mathrm{full}}$. Then, we increase w from zero (include only $\hat{M}$) to p (i.e., the model with no covariate, M_{null}, as the LBM, and the model with all of the covariates, M_{full}, as the UBM) so that (2.3) is "just" satisfied. As w increases, $\mathrm{NMCS}(w)$ gets "wider" and $\mathrm{CP}(w)$ gets closer to one. Therefore, there is a unique w such that $\mathrm{CP}(w-1) < 1-\alpha$ and $\mathrm{CP}(w) \geq 1-\alpha$. In the special case where $\mathrm{CP}(0) \geq 1-\alpha$, the NMCS reduces to a single model, $\hat{M}$. We can use the following bootstrap algorithm to obtain the final NMCS.

Algorithm 1: $100(1-\alpha)\%$ NMCS construction.

1. Generate B bootstrap samples $y^*_{[b]}$ under the distribution under model $\hat{M}$ and parameter vector $\hat{\psi}$.
2. For $b = 1, \ldots, B$, record the EO in lasso selection based on the bth bootstrap data $(y^*_{[b]}, X)$, denoted by $\mathrm{EO}^*_{[b]} = \{e^*_{1,[b]}, \ldots, e^*_{p,[b]}\}$, and the selected model $\hat{M}$, denoted by $\hat{M}^*_{[b]} = \{e^*_{1,[b]}, \ldots, e^*_{k_b,[b]}\}$.
3. For $w = 0, \ldots, 2p$, calculate the bootstrapped coverage probabilities of NMCS(w), $\mathrm{CP}^*(w) = \max_{0\le j\le w} B^{-1} \sum_{b=1}^{B} 1(\hat{M}^*_{-w+j,[b]} \subseteq \hat{M} \subseteq \hat{M}^*_{j,[b]})$, where $\hat{M}^*_{-w+j,[b]} = \{e^*_{1,[b]}, \ldots, e^*_{k_b-w+j,[b]}\}$, $\hat{M}^*_{j,[b]} = \{e^*_{1,[b]}, \ldots, e^*_{k_b+j,[b]}\}$, and $\hat{M}^*_x$ includes only the intercept term if $k_b + x \le 0$; $\hat{M}^*_x$ is the full model if $k_b + x \ge p$. Denote j that achieves the maximum as $f^*(w)$, which is a function of w.
4. Obtain the smallest w^* such that $\mathrm{CP}^*(w^*) \ge 1-\alpha$, and the corresponding $j^* = f^*(w^*)$.
5. The final $100(1-\alpha)\%$ NMCS is $(\hat{M}_{-w^*+j^*}, \hat{M}_{j^*})$, where $\hat{M}_{-w^*+j^*} = \{e_1, \ldots, e_{k-w^*+j^*}\}$ and $\hat{M}_{j^*} = \{e_1, \ldots, e_{k+j^*}\}$.

2.1.2. Extensions to GLMs and GAMs

Algorithm 1 is quite general, and thus can be extended to other models, such as GLMs and GAMs. In a GLM, the parameters in the distribution of y are connected to a linear predictor, $\eta = X\beta$, via a link function. Hence, we can select the variables and generate bootstrap samples from the distribution under model $\hat{M}$.

GAMs, introduced by Hastie and Tibshirani (1986), are generalized linear models in which the linear predictor can be represented as a sum of more general functions of a single variable: $\eta(X) = \sum_{j=1}^{p} f_j(X_j)$, where f_j are unknown functions, assumed to be smooth or otherwise have low complexity. Some model selection criteria delete irrelevant predictors or reduce the complexity of the f_j functions, such as the component selection and smoothing operator (COSSO) (Lin and Zhang (2006)), sparse additive model (SpAM) (Ravikumar et al. (2009)), the method of Meier, Geer and Buhlmann (2009), and generalized additive model selection (GAMSEL) (Chouldechova and Hastie (2015)). We adopt GAMSEL in our simulation study (see section 4.1.3); other selection methods can also be used. Specifically, GAMSEL represents f_j functions using a linear component and a nonlinear component; that is, $f_j(X_j) = \alpha_j X_j + \mu_j(X_j)^T \beta_j$, where μ_j is a vector of m_j Demmler–Reinsch basis (Demmler and Reinsch (1975)) functions. Then, GAMSEL generates sparse solutions by solving a convex optimization problem with L_1 penalties for α_j and group-lasso penalties for β_j, given by

$$\min_{\alpha_0,\alpha_j,\beta_j} \frac{1}{2}\Big\|y-\alpha_0-\sum_{j=1}^{p}\alpha_j X_j-\sum_{j=1}^{p}u_j(X_j)\beta_j\Big\|^2$$
$$+\lambda\sum_{j=1}^{p}\{\gamma|\alpha_j|+(1-\gamma)\|\beta_j\|_{D_j^*}\}+\frac{1}{2}\sum_{j=1}^{p}\psi_j\beta_j^T D_j\beta_j, \tag{2.7}$$

where $\|\beta_j\|_{D_j^*}=\sqrt{\beta_j^T D_j^*\beta_j}$, and D_j is the diagonal penalty matrix associated with the Demmler–Reinsch basis. The multiplier ψ_j is chosen to control the smoothness of the basis functions, that is, to achieve prespecified degrees of freedom when $\lambda=0$. The tuning parameter λ penalizes the linear and nonlinear coefficients of each term simultaneously, and sets all $\alpha_j=0$ and $\beta_j\equiv 0$ for large values. As λ decreases, some estimated coefficients become nonzero. If $\hat{\alpha}_j^\lambda$ and $\hat{\beta}_j^\lambda$ include nonzero elements, $\hat{f}_j^\lambda(X_j)$ is either a linear or a nonlinear nonzero function; hence, the variable X_j is selected into the model to construct $\eta(X)$ at the given λ. The EO in algorithm 1 can be obtained by recording the order of variables $\{X_1,\ldots,X_p\}$ that enter the model (when $\hat{f}_j(X_j)$ is nonzero) as λ decreases. The selected model $\hat{M}$ is the model at the cross-validated tuning parameter $\hat{\lambda}$. Then, we can generate the bootstrap samples from the distribution under $\hat{M}$ and the estimated parameters, and construct the NMCS using Algorithm 1.

2.2. LogP measure for $\hat{M}$

The LogP measure proposed by Liu, Li and Jiang (2021) quantifies the error in model selection. It is defined as the logarithm of the probability that the selected model is different from the optimal model,

$$\mathrm{LogP}=\mathrm{LogP}(\hat{M})=\log\{\mathrm{P}(\hat{M}\neq M_{\mathrm{opt}})\}. \tag{2.8}$$

The right side of (2.8) is evaluated using a bootstrapping procedure similar to Algorithm 1. First, we use $\hat{M}$ and $\hat{\psi}$ as approximations of M_{opt} and ψ_{opt}, respectively, to generate samples $y_{[b]}^*$, for $b=1,\ldots,B$, called bootstrap samples. Second, we perform model selection procedures for all $y_{[b]}^*$, and obtain the selected models, $\hat{M}_{[b]}^*$. Then, we calculate the empirical probability that $\hat{M}^*=\hat{M}$, and obtain the estimator of (2.8) as

$$\widehat{\mathrm{LogP}}=\log\left(1-\frac{1}{B}\sum_{b=1}^{B}1_{(\hat{M}_{[b]}^*=\hat{M})}\right). \tag{2.9}$$

Note that (2.9) is a natural estimator of LogP because of the following approximation supported by the law of large numbers:

$$\mathrm{P}(\hat{M}=M_{\mathrm{opt}})\approx\mathrm{P}(\hat{M}^*=\hat{M})\approx\frac{1}{B}\sum_{b=1}^{B}1_{(\hat{M}_{[b]}^*=\hat{M})}. \tag{2.10}$$

Here, we use the same LogP estimator, and extend the original information-based selection criteria to include shrinkage model selection methods. The consistency property of the LogP estimator and the conditions required to achieve these need to be re-examined. We also apply this estimator to more general classes of models, namely, GLMs and GAMs. The implementation is quite similar, because the LogP measure uses only the model selected from the data, $\hat{M}$ or $\hat{M}^*_{[b]}$, despite the different model structures and shrinkage methods.

3. Theoretical Properties

In this section, we study the theoretical properties of the proposed measures of uncertainty, including the coverage probability of the NMCS and the consistency of the LogP measure.

3.1. Coverage probability of the NMCS

Let y denote the original data, and $y^*_{[1]}, \ldots, y^*_{[B]}$ be the bootstrap samples. The notation $\mathrm{P}(\cdot|M, \psi)$ denotes the probability of the event when the underlying distribution and the parameter vector ψ are from model M. Let $\hat{\psi}_{\mathrm{opt}}$ denote the estimator of ψ_{opt} (i.e., $\hat{\psi}$ when $\hat{M} = M_{\mathrm{opt}}$). We make the following assumptions.

A1. The bootstrap samples $y^*_{[b]}$, for $1 \le b \le B$, are generated independently under $\hat{M}, \hat{\psi}$.

A2. There is a constant c such that, for every $w \ge 0$, $0 \le j \le w$, and fixed $\tilde{\psi}_{\mathrm{opt}}$, we have

$$\left|\mathrm{P}(M_{\mathrm{opt}} \in (\hat{M}^*_{-w+j}, \hat{M}^*_j)|M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) - \mathrm{P}(M_{\mathrm{opt}} \in (\hat{M}^*_{-w+j}, \hat{M}^*_j)|M_{\mathrm{opt}}, \psi_{\mathrm{opt}})\right| \le c|\tilde{\psi}_{\mathrm{opt}} - \psi_{\mathrm{opt}}|. \tag{3.1}$$

A3. $\mathrm{P}(\hat{M} = \hat{M}^*_l|\hat{M}, \hat{\psi}) > 0$, for every $l \ge 0$.

Note that the two probabilities inside the absolute value on the left side of (3.1) differ only in that the argument of ψ_{opt} in one probability is replaced by $\tilde{\psi}_{\mathrm{opt}}$ in the other. Furthermore, the absolute difference between the two probabilities is bounded by a constant times the absolute difference between ψ_{opt} and $\tilde{\psi}_{\mathrm{opt}}$. Therefore, this assumption is similar to a condition for Lipschitz continuity (e.g., Thomson, Bruckner and Bruckner (2008, p. 316)).

Proposition 1 (Linear regression with orthogonal covariates). *Let* (2.1.1) *hold, where* ε *has distribution* $N(0, \sigma^2 I_n)$*, with* σ^2 *a positive constant,* X *is an* $n \times p$ *matrix satisfying* $p = O(n^{c_1})$*, with* $0 \le c_1 < c_2 \le 1$*, and* $X^{\mathrm{T}}X = nI_p$*. The number of nonzero regression coefficients* q *is fixed, and* $n^{(1-c_2)/2} \min_{i=1,\ldots,q} |\beta_i| \ge M_0$*. For* λ *satisfying* $\lambda = O(\sqrt{n})$*, assumption A2 holds.*

A proof of Proposition 1 is given in the Supplementary Material.

Theorem 1. *Assume that assumptions A1–A3 hold. Let w^*, and j^* denote w and j, respectively, determined by step 4 in Algorithm 1. As $B \to \infty$, w^* and j^* converge in probability, with respect to the bootstrap distribution, to the integers $w \geq 0$ and $0 \leq j \leq w$, respectively, which depend on $\hat{M}$ and $\hat{\psi}$. Furthermore, for this w and j, we have*

$$\begin{aligned} &\mathrm{P}(M_{\mathrm{opt}} \in (\hat{M}_{-w+j}, \hat{M}_j)) \\ &\geq \frac{1-\alpha-\mathrm{P}(\hat{M} \neq M_{\mathrm{opt}}) - c\mathrm{E}\{|\hat{\psi}_{\mathrm{opt}} - \psi_{\mathrm{opt}}|1_{(\hat{M}=M_{\mathrm{opt}})}\} - o(1)}{\mathrm{P}(\hat{M} = M_{\mathrm{opt}})}, \end{aligned} \tag{3.2}$$

provided that $\mathrm{P}(\hat{M} = M_{\mathrm{opt}}) > 0$.

Theorem 1 establishes a lower bound for the coverage probability of the NMCS, (3.2), which depends on two quantities, namely, $\delta_1 = \mathrm{P}(\hat{M} \neq M_{\mathrm{opt}})$ and $\delta_2 = \mathrm{E}\{|\hat{\psi}_{\mathrm{opt}} - \psi_{\mathrm{opt}}|1_{(\hat{M}=M_{\mathrm{opt}})}\}$. Note that if $\hat{M}$ is a consistent model selector, then, under regularity conditions, we have $\delta_1 \to 0$, and the denominator of (3.2) goes to one, as the sample size, n, increases. Furthermore, if $\hat{\psi}_{\mathrm{opt}}$ converges in L^2 to ψ_{opt}, we have $\delta_2 \to 0$ as $n \to \infty$. Thus, as both $n, B \to \infty$, the limit of the right side of (3.2) is $1-\alpha$. The proof of Theorem 1 is given in the Supplementary Material.

3.2. Consistency in LogP estimation

Following Liu, Li and Jiang (2021), the consistency of $\widehat{\mathrm{LogP}}$ is defined as

$$\frac{\widehat{\mathrm{LogP}}}{\mathrm{LogP}} \xrightarrow{\mathrm{P}} 1, \tag{3.3}$$

as both $n, B \to \infty$, where the convergence in probability is with respect to the joint distribution of the data and the bootstrapping. The consistency holds under the following assumptions:

B1. $\hat{M}$ is consistent.

B2. $\mathrm{P}(\hat{M} = M_{\mathrm{opt}}) < 1$ and $\mathrm{P}(\hat{M}^* = M_{\mathrm{opt}}|M_{\mathrm{opt}}, \tilde{\psi}_{\mathrm{opt}}) < 1$, for any $\tilde{\psi}_{\mathrm{opt}}$.

B3. As $n \to \infty$, we have

$$\frac{\log\{\mathrm{P}(\hat{M}^* \neq M_{\mathrm{opt}}|M_{\mathrm{opt}}, \hat{\psi}_{\mathrm{opt}})\}}{\log\{\mathrm{P}(\hat{M}^* \neq M_{\mathrm{opt}}|M_{\mathrm{opt}}, \psi_{\mathrm{opt}})\}} \xrightarrow{\mathrm{P}} 1. \tag{3.4}$$

B4. Denote the probability in the numerator of (3.4) by P_*. For any $\eta > 0$, we have

$$\mathrm{E}\left[\frac{1-P_*}{P_*\{(1-P_*^{\eta}) \wedge (P_*^{-\eta}-1)\}^2}\right] < \infty. \tag{3.5}$$

Theorem 2 (Liu, Li and Jiang (2021)). *Under assumptions B1–B4,* (3.3) *holds in the sense that, for any $\epsilon > 0$ and $\rho > 0$, there are $N, B_n \geq 1$ that depend on ϵ, ρ, and B_n depends on n, such that*

$$\mathrm{P_J}\left(\left|\frac{\widehat{\mathrm{LogP}}}{\mathrm{LogP}} - 1\right| > \epsilon\right) < \rho, \quad n \geq N \text{ and } B \geq B_n, \tag{3.6}$$

where $\mathrm{P_J}$ denotes the joint probability of the data and the bootstrapping.

In the rest of this section, we establish the assumptions and Theorem 2 in a high-dimensional setting instead of the fixed–p case of Liu, Li and Jiang (2021). To further illustrate when the assumptions are met, we introduce the following notation. Let $\beta_{(1)} = (\beta_1, \ldots, \beta_q)$ be the nonzero coefficients in the true model. Let $X(1)$ and $X(2)$ be the first q and last $p - q$ columns of X, respectively, and $C = (1/n)X^TX$. By setting $C_{11} = (1/n)X(1)^TX(1)$, $C_{22} = (1/n)X(2)^TX(2)$, $C_{12} = (1/n)X(1)^TX(2)$, and $C_{21} = (1/n)X(2)^TX(1)$, C can be expressed in a block-wise form as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}. \tag{3.7}$$

Assume that C_{11} is nonsingular.

A strong irrepresentable condition (Zhao and Yu (2006)) is usually needed to establish the consistency of a lasso selection. Therefore, there exists a positive constant vector, η, such that

$$|C_{21}C_{11}^{-1}\mathrm{sign}(\beta_{(1)})| \leq \mathbf{1} - \eta, \tag{3.8}$$

where $\mathbf{1}$ is a $(p - q)$-dim vector of ones and the inequality holds element-wise.

In a high–dimensional model selection setting, we also need to constrain the dimension of the design matrix and the sparsity of the model parameters to establish consistency. We assume there exist $0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3, M_4 > 0$ such that the following hold:

$$\frac{1}{n}(X_i)'(X_i) \leq M_1, i = 1, \ldots, n, \tag{3.9}$$

$$\alpha' C_{11}\alpha \geq M_2, \text{ for } ||\alpha||_2^2 = 1, \tag{3.10}$$

$$q = O(n^{c_1}), \tag{3.11}$$

$$n^{(1-c_2)/2} \min_{i=1,\ldots,q} |\beta_i| \geq M_3, \tag{3.12}$$

Proposition 2 (Finite $2k$th moment). *Assume ε_i, for $i = 1, \ldots, n$, are i.i.d. random variables with a finite $2k$th moment,that is, $E(\varepsilon_i)^{2k} < \infty$, for an integer $k > 0$, and the design matrix satisfies the strong irrepresentable condition and* (3.9)–(3.12). *For $p = o(n^{(c_2-c_1)k})$, and any λ that satisfies $\lambda/\sqrt{n} = o(n^{(c_2-c_1)/2})$ and $p^{-1}(\lambda/\sqrt{n})^{2k} \to \infty$, assumptions B1–B4 hold.*

A proof of Proposition 2 can be found in the Supplementary Material. Note that the conditions in Proposition 2 are not more restrictive than those assumed for sign consistency (Zhao and Yu (2006)), which suggests that assumptions B1–B4 can be achieved easily under general settings. In particular, for Gaussian noise that has all the moments, the growing rate of p can be relaxed to an exponential rate for selection consistency. Hence, it is possible for assumptions B1–B4 to be satisfied with exponentially growing p under Gaussian noise. This yields the following result.

Proposition 3 (Gaussian noise). *Assume ε_i, for $i = 1, \ldots, n$, are i.i.d. Gaussian random variables, and that the design matrix satisfies the strong irrepresentable condition and* (3.9)-(3.12). *If there exist $0 \le c_3 < c_2 - c_1$ such that $p = O(e^{n^{c_3}})$, for $\lambda \propto n^{(1+c_4)/2}$ with $c_3 < c_4 < c_2 - c_1$, assumptions 1–B4 hold.*

4. Simulation Studies

4.1. Simulation studies for the NMCS

We investigate the performance of the NMCS under three types of models: linear regression models, logistic regression models, and GAMs.

As shown in the theorems in Section 3, the sign consistency of the lasso and the theoretical properties of the NMCS and LogP are usually based on $p = o(n^c)$, for some $0 < c < 1$, for general design matrices and noise distributions. We first run a screening procedure called iterative sure independence screening to reduce the dimensionality of p from a possibly huge scale [say, $\exp\{O(n^\delta)\}$, for some $\delta > 0$] to a scale that is more manageable [e.g., $o(n)$] via a fast and efficient algorithm; see Fan and Li (2008) and Fan and Song (2010). The algorithm is implemented using the R package *SIS*. We then construct MCSs using the NMCS and its competitor based on the surviving features. We consider $n = 200$, $p = 1000$, and $d = n - 1 = 199$ for the linear and GLM simulations.

4.1.1. Linear regression

For a linear regression, we compare the performance of the NMCS and the MCB method of Li et al. (2019) under the same simulation setting.

The covariates are generated under one of the following cases:

Case 1: $x_1, \ldots, x_p$ are i.i.d. N(0, 1) random variables.

Case 2: $x_1, \ldots, x_p$ are jointly Gaussian, marginally distributed as N(0, 1), and have the correlation structure $\mathrm{cor}(x_i, x_j) = 0.5^{|i-j|}$.

Case 3: $x_1, \ldots, x_p$ are jointly Gaussian, marginally distributed as N(0, 1), and have the correlation structure $\mathrm{cor}(x_i, x_j) = 0.5$, for $i \neq j$.

Case 1 is a simple case for variable selection. Case 2 has an exponential decay correlation structure between the predictors, and Case 3 has a constant

correlation as $|i-j|$ increases. These simulation cases are adapted from Li et al. (2019), Fan and Li (2008), and Fan, Yang and Song (2011).

The true parameters are such that β_1, β_2, and β_3 are generated randomly as $(5\log n/\sqrt{n}+|Z|)U$, with $Z \sim N(0,1)$, U independent of Z with $\mathrm{P}(U=1)=0.4$ and $\mathrm{P}(U=-1)=0.6$, and $\beta_j = 0$ for $j > 3$.

After the covariates are generated, we generate the responses using $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, where $i = 1, \ldots, n$ and $\epsilon_i \sim N(0,1)$. The NMCS and MCB are both based on a BIC-tuned adaptive lasso (ALasso) and $B = 500$ bootstrap samples. We also record the simulated coverage probability and mean width of the MCSs, and use P^* to denote the empirical probability (based on the simulation runs) that $\hat{M} = M_{\mathrm{opt}}$. The results are shown in Table 1 in the Supplementary Material.

In all simulation runs, the true predictors x_1, x_2, x_3 were successfully selected by the initial screening. We compare the NMCS with the MCB method of Li et al. (2019) in terms of both the empirical coverage probability (CP) and average width (AW) of the MCSs. The CP for both methods are very similar in all cases, and are higher than the nominal confidence level (CL). Comparing AW, the NMCS has a smaller width when the CL is large or the covariates are correlated (under Cases 2 and 3, respectively), which suggests that the NMCS is more efficient in these cases. Furthermore, the NMCS is more stable than the MCB in terms of keeping smaller values of AW in all cases. For example, when the AW of the MCB is larger than 10, the NMCS has a significant advantage. Note that the construction of the MCB uses only the single selected model, whereas the NMCS records and uses all the models in the full solution path of the shrinkage model selection, which explains the advantage of the NMCS.

4.1.2. Logistic regression

We consider covariates generated under Cases 1, 2, and 3 described in the previous subsection. We then simulate the response variable, Y, from a Bernoulli distribution with the probability of success $p(x)$) such that $\log[p(x)/\{1-p(x)\}] = x'\beta$. In this case, the NMCS is constructed using the BIC-tuned ALasso with $B = 500$ bootstrap samples.

The results, based on $K = 200$ simulation runs, are presented in Table 2. Here, the initial screening did not fully preserve all of the true predictors under Cases 2 and 3. Therefore, we record the frequency when the screening preserves all of the true predictors in $K = 200$ simulation runs as "SIS_rate." Unlike in the linear case, there are no existing methods for a comparison. The CP is above the nominal CL under Cases 1 and 2. Under Case 3, the CP of the 95% NMCS is lower than 0.95; however, in all runs in which the initial screening protects all the true predictors, the "conditional CP"$(0.905/0.970 = 0.933)$ meets the nominal CL. Furthermore, the accuracy of the selected model, P^*, is not close to one in any of the three cases, but the NMCS still performs satisfactory, showing that the

performance of the NMCS is not sensitive to the accuracy of the selected model, as suggested by Theorem 1. In addition, the AW decreases with the CL, which is reasonable.

4.1.3. GAM

In this case, the covariates are generated under the following setting:

Case 4: $\{x_k\}_{k\neq 2}$ are i.i.d. N(0, 1) random variables, where $x_2 = -(1/3)x_1^3 + \tilde{\epsilon}$ and $\tilde{\epsilon} \sim N(0,1)$. The true parameters are $\beta_1 = \beta_2 = \beta_3 = 1$ and $\beta_j = 0$, for $j > 3$.

The responses are generated under the model $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon$, where $\epsilon \sim N(0,1)$ and is independent of $\tilde{\epsilon}$. Note that, in this case, $\mathrm{E}(Y|x_1)$ and $\mathrm{E}(Y|x_2)$ are nonlinear about x_1 and x_2, respectively. We use GAMSEL (Chouldechova and Hastie (2015)) with $m_j = 6$ basis functions and three degrees of freedom as the GAM selection method, which can be implemented using the R package *gamsel*. Because GAMSEL estimates significantly more parameters than the GLM does, the consistency of GAMSEL is more difficult to establish in the ultra-high-dimensional case. Hence, we consider a relatively high–dimensional case with $p = 30$ for the GAM simulation, without initial screening. The bootstrap sample size is $B = 400$.

The results, based on $K = 100$ simulation runs, are presented in Table 3. Again, there are no existing methods available for a comparison, but the CP meets the nominal CL; the AW decreases with the CL, but the scale of the decrease is much smaller than those observed in Tables 1 and 2.

4.2. Simulation study for LogP

The LogP measure is conceptually and computationally easier under all three types of models. Thus, as an example, we present simulation results on the performance of LogP for a linear model selection under Case 1 of Section 4.1.1. We consider $\delta = 10^{-4}$, $n = (50, 100, 200, 500)$, $p = \lfloor n^{3/4} \rfloor$, and $B = (500, 1000)$.

A performance measure for LogP estimation is the percentage relative bias, defined as $\%\mathrm{RB} = 100 \times [\{\mathrm{E}(\widehat{\mathrm{LogP}}) - \mathrm{LogP}\}/\mathrm{LogP}]$, where LogP is the simulated true LogP, that is, the logarithm of the empirical probability, based on the simulation runs, that $\hat{M} \neq M_{\mathrm{opt}}$, and $\mathrm{E}(\widehat{\mathrm{LogP}})$ is the mean of the LogP estimator, also based on the simulation runs. Another performance measure is the coefficient of variation, defined as $\mathrm{CV} = \mathrm{s.d.}(\widehat{\mathrm{LogP}})/|\mathrm{E}(\widehat{\mathrm{LogP}})|$, where $\mathrm{s.d.}(\widehat{\mathrm{LogP}}) = \{\mathrm{var}(\widehat{\mathrm{LogP}})\}^{1/2}$ and $\mathrm{var}(\widehat{\mathrm{LogP}})$ is the variance of $\widehat{\mathrm{LogP}}$ based on the simulation runs.

We consider LogP for BIC-tuned lasso, ALasso, and SCAD selections. The mean of the estimated LogP and the corresponding RB% and CV, based on $K = 200$ simulation runs, are presented in Table 4, where P^* denotes the empirical probability (based on the simulation runs) that $\hat{M} = M_{\mathrm{opt}}$.

In terms of the %RB, the SCAD performs satisfactorily. In general, the performance in terms of %RB is considered satisfactory if it is a single-digit or low double-digit number (10s or 20s). The SCAD and ALasso perform satisfactorily when p is relatively small (< 100). However, the performance of the lasso is not satisfactory. This may be because the SCAD and ALasso provide more accurate estimators of the model parameters. It is known (Fan and Li (2001); Zou (2006)) that SCAD and ALasso estimators have the oracle property, whereas the lasso estimator is consistent under more restrictive conditions.

In addition, for ALasso, the performance of $\widehat{\text{LogP}}$ is satisfactory, as long as P^* is not very close to one. As noted in Liu, Li and Jiang (2021), the LogP measure is more useful when $\text{P}(\hat{M} = M_{\text{opt}})$, that is, the probability of choosing the optimal model using the model selector, is not very close to one (because in the latter case, there is not much uncertainty associated with the model selection). Thus, practically, the performance of $\widehat{\text{LogP}}$ for ALasso is considered satisfactory for the cases that matter.

Overall, the simulation results show some interesting differences between the three most popular shrinkage selection/estimation procedures, namely, the lasso, ALasso, and SCAD, that seemingly favor the ALasso and SCAD.

5. Real–Data Examples

We provide two real-data examples to illustrate the proposed measures of uncertainty for model selection. The first example is under a linear model setting, and the second is under a logistic regression framework.

5.1. NDD1 data analysis

Conlon et al. (2003) discussed a motif regression problem. The objective is to find binding sites in DNA sequences of an NDD1 transcriptional activator (TA) that is essential for the expression of a set of late-S-phase-specific genes. The binding sites are called motifs and are short sequences of the DNA codes A, C, G, and T. The number of candidate motifs is $p = 100$, and the number of DNA segments in the data is $n = 66$. The response y is the measurement of the binding intensity of the NDD1 activator on the DNA segments. The variable x_j is a measure of the abundance score of candidate motif j in the DNA segment. We can fit a linear regression model using y and x_j. We first run sure independence screening (Fan and Li (2008)) to reduce the dimension to $d = n - 1 = 65$, and then compute the NMCS, MCB (Li et al. (2019)), and LogP measures using the surviving predictors.

Here, as model selection methods, we use ALasso with the tuning parameter chosen using the BIC (ALasso-BIC) or 10-fold cross-validation (ALasso-CV), and select the most parsimonious model within one standard error of the minimum (Hastie, Tibshirani and Friedman (2001)). The results are presented in Table 5,

and show that although the selected model $\hat{M}$ varies depending on the tuning methods, almost all of the 95% confidence intervals contain all of the models selected using the various methods. Here, NULL means an empty set, that is, no predictor is selected. Comparing the widths of the confidence sets, the NMCS is more efficient than the MCB at the chosen CLs. The model with predictors {1, 4, 5, 80} is included in almost all confidence sets, suggesting that these motifs may be the binding sites of the TA. Other motifs in the UBM of the 95% NMCS cannot be excluded at the 95% CL, hence, additional data are required to further investigate the effects of these motifs. These results support those of Pang, Lin and Jiang (2016), who found that the first and fourth motifs contain the consensus pattern from the Saccharomyces Genome Database (Chen et al. (2008)).

5.2. South African heart disease data

Hastie, Tibshirani and Friedman (2001) considered a data set on heart disease in South African men. The data include 462 observations. Some potential predictors are considered and indexed by $1, \ldots, 9$, indicating systolic blood pressure (`sbp`), cumulative tobacco (in kg; `tobacco`), low density lipoprotein cholesterol (`ldl`), amount of fat found in adipose tissue (`adiposity`), family history of heart disease (`famhist`), type-A behavior rating (`typea`), obesity score (`obesity`), current alcohol consumption (`alcohol`), and age at onset (`age`), respectively. The responses are binary indicators of whether or not the person had heart disease.

We apply the NMCS and the LogP measures to those data. Available methods for a shrinkage selection/estimation under a GLM setting include the lasso and ALasso (see Section 4.1.2). The results are presented in Table 6. Although the selected model, $\hat{M}$, varies depending on the selection criteria, almost all of the 95% confidence sets contain all of the selected models. For example, the 95% NMCS of the AIC-tuned lasso includes all of the models selected by any of the six methods, that is, the lasso or ALasso tuned using the AIC, BIC, or CV. The widths of the NMCS and the values of LogP indicate the uncertainty of different selection methods. The smaller LogP of the BIC-tuned ALasso suggests that its selected model, $\{2, 3, 5, 6, 9\}$, is more likely to be the optimal model. The narrower NMCS of the CV-tuned ALasso suggests there is less variation among the selected models. Furthermore, the ALasso methods have smaller errors and less variation than the lasso methods do. Thus, we may favor a model selected using BIC-tuned ALasso or CV-tuned ALasso, depending on whether we need to reduce the error or the variation. The model with predictors {`tobacco`(2), `ldl`(3), `famhist`(5), `age`(9)} is included in all confidence sets, indicating the importance of these predictors in terms of explaining heart disease. The importance of the predictors `sbp`(1), `typea`(6), and `obesity`(7) requires further investigation. These conclusions are consistent with those of Hastie, Tibshirani and Friedman (2001), in which the model $\{2, 3, 5, 9\}$ is selected under a GLM setting, and `sbp`(1) and

`obesity`(7) have nonlinear effects that can be added to the model under a GAM setting.

6. Conclusion

We have proposed two measures of uncertainty for shrinkage model selection procedures that can be used in high–dimensional cases (p grows at a polynomial/exponential rate) and for more general classes of models. The NMCS is obtained by trimming the EO of the variables in the solution paths when the tuning parameter decreases. We prove that, under some general conditions, the optimal model is contained in the NMCS at least at a probability that is close to the nominal CL, if the model selection method and the parameter estimators are consistent. Our simulations show that the empirical coverage probabilities of the NMCS meet the nominal CLs even when the covariates are highly correlated and p is much larger than n (decreased to $n-1$ by using initial screening). Compared with other MCS methods, such as the MAC (Liu, Li and Jiang (2021)) and MCB (Li et al. (2019)), the NMCS has advantages in terms of applicability and efficiency. Another measure, LogP, measures the error of a (single) selected model by estimating the logarithm of the probability that the selected model is different from the optimal model. The LogP estimator based on bootstrap probabilities is consistent under some weak conditions. The results of simulation studies show that the LogP estimator has satisfactory accuracy when the selected model exhibits relatively large to modest uncertainty (the probability that the selected model is the true model is between 0.5 and 0.98; see Table 4).

The proposed measures can be used to signal the uncertainty of a selected model by checking the width of the NMCS (a larger width means higher "variance") and the value of LogP (a smaller value means less "bias"). A comparison of these measures for different model selection methods using simulation data and real data shows that the ALasso and SCAD penalty generate selected models with less "bias" than those selected using the lasso penalty. In addition, the selected models tuned using the cross-validation with the one standard error rule have lower "variance" than those tuned using AIC/BIC. These findings are useful when considering which selection method to use, and even whether model selection makes sense, given the level of uncertainty associated with the data.

Furthermore, the nested property of the NMCS provides sufficient options for choosing a model, from relatively parsimonious models to more conservative and higher-dimensional models, with a given CL. In some areas with expensive data collection, such as experimentation data, the NMCS tells us which of the covariates are most important covariates (those in the LBM), as well as identifying some potential important (those in the UBM but not included in the LBM) with a probability guarantee. Here, potential applications include speeding up the

drug development process and helping to allocate a budget efficiently.

Supplementary Material

A web appendix, referenced throughout the manuscript, contains technical proofs and tables and is available online.

Acknowledgments

The research of Jiming Jiang was supported by the NSF of the United States, grants DMS-1713120 and DMS-1914465. The authors are grateful to the associate editor and two referees for their helpful comments and suggestions.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267–281. Akademiai Kiado, Budapest.

Chen, X., Guo, L., Fan, Z. and Jiang, T. (2008). W-AlignACE: An improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics* **24**, 1121–1128.

Chouldechova, A. and Hastie, T. (2015). Generalized additive model selection. *arXiv 1506.03850.*

Conlon, E., Liu, X., Lieb, J. and Liu, J. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS* **100**, 3339–3344.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik* **24**, 375–382.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–499.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129.

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

Fan, J. and Li, J.(2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70**, 849–911.

Fan J and Song R (2010). Sure independence screening in generalized linear models with NP dimensionality. *Ann. Statist.* **38**, 3567–3604.

Fan, J., Yang, F. and Song, R.(2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557.

Ferrari, D. and Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statist. Sinica* **25**, 1637–1658.

Fu, W. J., Carroll, R. J. and Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21**, 1979–1986.

Hansen, P. R., Lunde, A. and Nason, J. M. (2011). The model confidence set. *Econometrica* **79**, 453–497.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.* **1**, 297–310.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

Jiang, J. and Nguyen, T. (2016). *The Fence Methods*. World Scientific, Singapore.

Li, Y., Luo, Y., Ferrari, D., Hu, X. and Qin, Y. (2019). Model Confidence Bounds for Variable Selection. *Biometrics* **75**, 392–403.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist* **34**, 2272–2297.

Liu, X., Li, Y. and Jiang, J. (2021). Simple measures of uncertainty for model selection. *TEST* **30**, 1–20.

Lubke, G. H. and Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. *Struct. Equ. Model.* **23**, 479–490.

Lubke, G. J., Campbell, I., McArtor, D., Miller, P., Luningham, J. and van den Berg, S. M. (2017). Assessing model selection uncertainty using a bootstrap approach: An update. *Struct. Equ. Model.* **24**, 230–245.

Meier, L., Geer, S., Buhlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779–3821.

Pang, Z., Lin, B. and Jiang, J. (2016). Regularisation parameter selection via bootstrapping. *Aust. N. Z. J. Stat.* **58**, 335–356.

Ravikumar, P. D., Liu, H., Lafferty, J. D. and Wasserman, L. A. (2009). SpAM: Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 1009–1030

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

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**, 15545–15550.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **16**, 385—395.

Thomson, B. S., Bruckner, J. B. and Bruckner, A. M. (2008). *Elementary Real Analysis*. 2nd Edition. Prentice-Hall.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 671–683.

Zhao, P. and Yu, B (2006). On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Zheng, C., Ferrari, D. and Yang, Y. (2019). Model selection confidence sets by likelihood ratio testing. *Statist. Sinica* **29**, 827–851.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Yuanyuan Li

Department of Statistics, University of California, Davis.

E-mail: yynli9696@gmail.com

Jiming Jiang

Department of Statistics, University of California, Davis.

E-mail: jimjiang@ucdavis.edu

(Received August 2021; accepted July 2022)