Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Parsimonious Model Averaging With a Diverging Number of Parameters

Xinyu Zhang, Guohua Zou, Hua Liang & Raymond J. Carroll

To cite this article: Xinyu Zhang, Guohua Zou, Hua Liang & Raymond J. Carroll (2020) Parsimonious Model Averaging With a Diverging Number of Parameters, Journal of the American Statistical Association, 115:530, 972-984, DOI: 10.1080/01621459.2019.1604363

To link to this article: https://doi.org/10.1080/01621459.2019.1604363

Accepted author version posted online: 19 Apr 2019.
Published online: 19 Jun 2019.

Submit your article to this journal

Article views: 600

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 530, 972–984: Theory and Methods
https://doi.org/10.1080/01621459.2019.1604363

Taylor & Francis
Taylor & Francis Group

Check for updates

Parsimonious Model Averaging With a Diverging Number of Parameters

Xinyu Zhang[a,b], Guohua Zou[c], Hua Liang[d], and Raymond J. Carroll[e]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; [b]School of Mathematics and Statistics, Qingdao University, Qingdao, China; [c]School of Mathematical Sciences, Capital Normal University, Beijing, China; [d]Department of Statistics, George Washington University, Washington, DC; [e]Department of Statistics, Texas A&M University and University of Technology Sydney

ABSTRACT

Model averaging generally provides better predictions than model selection, but the existing model averaging methods cannot lead to parsimonious models. Parsimony is an especially important property when the number of parameters is large. To achieve a parsimonious model averaging coefficient estimator, we suggest a novel criterion for choosing weights. Asymptotic properties are derived in two practical scenarios: (i) one or more correct models exist in the candidate model set and (ii) all candidate models are misspecified. Under the former scenario, it is proved that our method can put the weight one to the smallest correct model and the resulting model averaging estimators of coefficients have many zeros and thus lead to a parsimonious model. The asymptotic distribution of the estimators is also provided. Under the latter scenario, prediction is mainly focused on and we prove that the proposed procedure is asymptotically optimal in the sense that its squared prediction loss and risk are asymptotically identical to those of the best—but infeasible—model averaging estimator. Numerical analysis shows the promise of the proposed procedure over existing model averaging and selection methods.

ARTICLE HISTORY
Received August 2017
Revised February 2019

KEYWORDS
Asymptotic optimality; Frequentist model averaging; Jackknife model averaging; Mallows model averaging; Parsimony.

1. Introduction

In the past two decades, there has been considerable research devoted to model averaging; see Buckland, Burnham, and Augustin (1997), Hjort and Claeskens (2003), Hansen (2007), Zhang and Liang (2011), Liu and Okui (2013), Zhang, Zou, and Liang (2014), and Lu and Su (2015), among others. Model averaging has long been popular among Bayesian statisticians. Reviews of the relevant Bayesian literature can be found in Hoeting et al. (1999). In this article, we focus only on frequentist model averaging. A primary motivation of model averaging is that it often reduces the prediction risk in regression estimation, as "betting" on multiple models provides a type of insurance against a singly selected model being poor (Leung and Barron 2006). In particular, model averaging usually improves estimation accuracy when the underlying model is unstable with a high noise level (Yuan and Yang 2005; Zhang, Wan, and Zhou 2012). Additionally, it is often the case that several models fit the data equally well, but may differ substantially in terms of the variables included and may lead to very different predictions (Miller 2002). Thus, combining these models is more reasonable than choosing just one of them.

This article develops a specific model averaging procedure inspired by the following four aspects.

(a) (*Diverging* p_n) Many existing model averaging methods were developed under a fixed p_n, including the smoothed focused information criterion weighting strategy of Hjort and Claeskens (2003) and the optimal model averaging method of Liang et al. (2011). Their theoretical properties may not hold when p_n increases with the sample size n, which is commonly in a high-dimensional setting. Whether similar criteria can be developed and whether associated theoretical properties still hold under the diverging p_n scenario remain unclear and not examined in the literature, at least to the best of our knowledge. Our method will be developed under the diverging p_n scenario.

(b) (*Parsimony*) Parsimony is an especially important property when modeling a dataset with a diverging dimension. Mallows model averaging (MMA) (Hansen 2007) and Jackknife model averaging (JMA) (Hansen and Racine 2012) were developed with nonfixed p_n, and the latter was recently extended to high-dimensional settings by Ando and Li (2014). However, the number of nonzero coefficient estimators by MMA or JMA method is large for high-dimensional regression because it can put substantial weights on overfitted models, as found in our numerical examples. We investigate the parsimony property for our model averaging procedure, showing that it adaptively puts zero weights on overfitted models.

(c) (*Asymptotic distribution*) Although many model averaging methods have been developed, little attention has been given to the study of the asymptotic distribution of the resulting model averaging estimator (e.g., in Ando and Li (2014), the asymptotic distribution of their model averaging estimator was not studied). Exceptions are Hjort and Claeskens (2003) and literature following their work, such as Zhang and Liang (2011) and Liu (2015), where asymptotic distribution theories are built under a local misspecification framework in which some coefficients are of order $n^{-1/2}$, here n is the sample size. Although this framework provides a useful

CONTACT Hua Liang hliang@gwu.edu Department of Statistics, George Washington University, 801 22nd St. NW, Washington, DC 20052.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
© 2019 American Statistical Association

tool for the study of the asymptotic distributions of the model averaging estimators, there are some limitations on this framework (see Ishwaran and Rao 2003; Raftery and Zheng 2003). In addition, the asymptotic distributions are nonnormal and have complicated forms, which makes the inference difficult. Without using the local misspecification framework, the asymptotic normality of our model averaging estimator will be derived. This makes the inference feasible.

(d) (*Robustness for misspecification*) Asymptotic optimality in the sense of minimizing squared prediction loss is built in the existing literature such as Wan, Zhang, and Zou (2010) and Liu and Okui (2013), where, but, all candidate models are essentially needed to be misspecified. Typically, when focusing on coefficient estimation or model selection, people consider the scenario that there are correct models in candidate set, under which, the existing literature has no theoretical support for their model averaging. Our model averaging estimator is robust to either there are correct models in candidate set or all candidate models are misspecified, and our method has theoretical justifications under both two scenarios. Under the former one, our method puts the weight one to the smallest correct model and leads to a parsimonious model. The asymptotic distribution of the estimators is also provided. Under the latter one, prediction is mainly focused on and we prove that the proposed procedure is asymptotically optimal in the sense that its squared prediction loss and risk are asymptotically identical to those of the best—but infeasible—model averaging estimator.

The remainder of this article is organized as follows. Section 2 proposes the model averaging estimation procedure and presents its theoretical properties. Section 3 investigates its finite sample performance by numerical examples. Section 4 provides some discussions. Proofs of the results are presented in the Appendix.

2. Estimation and Main Results

Consider the model

$$y_i = \sum_{m=1}^{p_n} x_{im}\beta_m + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with mean 0 and variance σ^2, $x_{i1}, \ldots, x_{ip_n}$ are nonstochastic predictors, and p_n is the number of predictors. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^{\mathrm{T}}$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, and the predictors $\mathbf{x}_m = (x_{1m}, \ldots, x_{nm})^{\mathrm{T}}$, $m = 1, \ldots, p_n$. Write the predictor matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{p_n}]$. *Note that p_n is diverging, that is, it can increase when the sample size n increases.* Let $\widetilde{\mathcal{A}} = \{m : \beta_m \neq 0\}$. The cardinality of $\widetilde{\mathcal{A}}$, p_0, *can also be diverging*. Let $\widetilde{\mathbf{X}}^*$ be the $n \times p_0$ matrix composed of predictors with all nonzero coefficients and the model with $\widetilde{\mathbf{X}}^*$ being its regressor matrix is called *true* model. We focus on the case $p_n < n$ in Sections 2.1–2.5, and extend to the case $p_n > n$ in Section 2.6.

2.1. Candidate Models and Weight Choice Criterion

Without any constraints, there are 2^{p_n} candidate models, which is huge when p_n is large or even just moderate, say 50. Applying model averaging methods in such a case is computationally infeasible. Thus, prior to model averaging, we have to prepare candidate models. Following some model averaging literature, such as Hansen (2007) and Hansen (2014), we use nested candidate models, that is, the sth candidate model using the first v_s predictors, such that $0 < v_1 < v_2 < \cdots < v_{q_n}$, where q_n is the number of candidate models. Therefore, $q_n \ll 2^{p_n}$ and thus the computation burden will be largely reduced. In Section 2.5, we will show how to prepare nested candidate models.

Let $\mathbf{X}_j$ be the predictor matrix in the jth candidate model and $\boldsymbol{\Pi}_j$ be a selection matrix such that $\mathbf{X}\boldsymbol{\Pi}_j = \mathbf{X}_j$. We assume $\mathbf{X}_j$ be of full column rank, then the estimate of $\boldsymbol{\beta}$ under the jth candidate model is $\widehat{\boldsymbol{\beta}}_j = \boldsymbol{\Pi}_j(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j)^{-1}\mathbf{X}_j^{\mathrm{T}}\mathbf{y}$. Let the weight vector $\mathbf{w} = (w_1, ..., w_{q_n})^{\mathrm{T}}$ belong to the set $\mathcal{W} = \{\mathbf{w} \in [0,1]^{q_n} : \sum_{j=1}^{q_n} w_j = 1\}$. The average estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{j=1}^{q_n} w_j\widehat{\boldsymbol{\beta}}_j$. We further select the weight vector by minimizing the criterion

$$\mathcal{S}_n(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w})\|^2 + \phi_n\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v}, \tag{2}$$

that is, $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \mathcal{S}_n(\mathbf{w})$, where $\mathbf{v} = (v_1, \ldots, v_{q_n})^{\mathrm{T}}$, v_j is the number of columns of $\mathbf{X}_j$, $\widehat{\sigma}^2 = (n - v_{j^*})^{-1}\|\mathbf{y} - \mathbf{X}_{j^*}\widehat{\boldsymbol{\beta}}_{j^*}\|^2$ is an estimator of σ^2, j^* is the index of a submodel, and ϕ_n is a positive scale depending on n. The estimator of $\widehat{\sigma}^2$ and the choice of ϕ_n will be discussed in Section 2.5. When $\phi_n = 2$, the criterion $\mathcal{S}_n(\mathbf{w})$ corresponds the Mallows criterion of Hansen (2007).

The model averaging estimator of $\boldsymbol{\beta}$ is defined as $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) = \sum_{j=1}^{q_n} \widehat{w}_j\widehat{\boldsymbol{\beta}}_j$, where $\widehat{w}_j$ is the jth element of $\widehat{\mathbf{w}}$. Because $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ will be proved to have the parsimony property under some regularity conditions, we term our model averaging method the parsimonious model averaging (PMA).

2.2. Parsimony Under the Scenario That There Are Correct Candidate Models

The candidate models including $\widetilde{\mathbf{X}}^*$ are defined as *correct* models. Let q^* be the index of *the smallest correct model*, whose predictor matrix is denoted by $\mathbf{X}^*$. Let $\mathcal{A} = \{m \in \{1, \ldots, p_n\} : \mathbf{x}_m \text{ is a column of } \mathbf{X}^*\}$. Note that $\widetilde{\mathcal{A}} \subseteq \mathcal{A}$. Let $\mathcal{O}$ be a set including all overfitted candidate models, that is, for any $j \in \mathcal{O}$, $\mathbf{X}_j$ includes all columns of $\mathbf{X}^*$, but $\mathbf{X}_j \neq \mathbf{X}^*$. Unless otherwise stated, all limiting processes discussed in this and subsequent sections are as $n \to \infty$. To build the parsimony, we need the following condition.

Condition (C.1). There exist positive constants $0 < c_1 < c_2 < \infty$ such that

$$\Pr(c_2 \geq \widehat{\sigma}^2\sigma^{-2} \geq c_1) \to 1.$$

Condition (C.1) requires that $\widehat{\sigma}^2$ does not approach zero or infinity, as $n \to \infty$ and as implied by condition 2 of Yuan and Yang (2005). Note that Condition (C.1) does not require that $\widehat{\sigma}^2$ is consistent and thus is very easily satisfied.

Lemma 1. If Condition (C.1) is satisfied, and $\phi_n \to \infty$, then

$$\Pr(\widehat{w}_j = 0 \text{ for all } j \in \mathcal{O}) \to 1. \tag{3}$$

Remark 1. In light of $\phi_n \to \infty$, our criterion $\mathcal{S}_n(\mathbf{w})$ is different from the Mallows criterion of Hansen (2007), in which $\phi_n = 2$.

Lemma 1 means that when there are correct candidate models, zero weights are put on overfitted candidate models (i.e., the models larger than the smallest correct model) with probability approaching 1. Let $\widehat{\beta}(\widehat{\mathbf{w}})_m$ be the mth element of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ and $\mathcal{A}^c$ be the complement of $\mathcal{A}$. Lemma 1 implies the following result.

Theorem 1 (Parsimony). Under the conditions of Lemma 1,

$$\Pr\{\widehat{\beta}(\widehat{\mathbf{w}})_m = 0 \text{ for all } m \in \mathcal{A}^c\} \to 1. \tag{4}$$

2.3. Asymptotic Distribution Under the Scenario That There Are Correct Candidate Models

Let $\mathcal{U}$ be a set including all underfitted candidate models, that is, for any $j \in \mathcal{U}$, $\mathbf{X}^*$ includes all columns of $\mathbf{X}_j$ but $v_j < v_{q^*}$. Let $\lambda_{\min}(\mathcal{M})$ and $\lambda_{\max}(\mathcal{M})$ be the minimum and maximum eigenvalues of a positive definite matrix $\mathcal{M}$, respectively.

Condition (C.2). There exists a positive constant κ_1 such that $\kappa_1 \le \lambda_{\min}(n^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)$.

Lemma 2. Under Conditions (C.1)–(C.2), if $\phi_n \to \infty$ and $\phi_{nv_{q^*}} n^{-1} \to 0$, then for any $j \in \mathcal{U}$,

$$\widehat{w}_j = O_p(\phi_n v_{q^*} n^{-1}). \tag{5}$$

Lemma 2 means that the weight on any underfitted candidate model is bounded by $\phi_n v_{q^*} n^{-1}$ in probability and will be used in the proof of Theorem 2. Combining Lemmas 1 and 2, we reach a very important conclusion as follows.

> *When ϕ_n is set appropriately and the true model is in the candidate set, the weight on the true model converges to one as $n \to \infty$. If the true model is not in the candidate set but there are overfitted models in the candidate set, then from the proofs of Lemmas 1 and 2, it is straightforward to obtain that the weight on the smallest correct model converges to one as $n \to \infty$.*

Based on the above theoretical results, we now build the oracle property of our averaging method, including variable selection consistency and asymptotic normality. Let $\boldsymbol{\beta}_{\mathcal{A}}$ be the vector of the coefficient vector in the smallest correct model q^*, $\widehat{\boldsymbol{\beta}}_{\mathcal{A},j}$ be the estimate of $\boldsymbol{\beta}_{\mathcal{A}}$ under the jth candidate model, the averaging estimate of $\boldsymbol{\beta}_{\mathcal{A}}$ be $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})_{\mathcal{A}} = \sum_{j=1}^{q_n} \widehat{w}_j \widehat{\boldsymbol{\beta}}_{\mathcal{A},j}$, $\widehat{\beta}(\widehat{\mathbf{w}})_{\mathcal{A},m}$ be the mth element of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})_{\mathcal{A}}$, and $\widehat{\mathcal{A}} = \{m \in \{1,\ldots,p_n\} : \widehat{\beta}(\widehat{\mathbf{w}})_m \neq 0\}$.

Condition (C.3). $n^{-1}\max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{v_{q^*}} x_{ij}^2 = o(1)$, and for some constants $\kappa_2 > 0$, $\kappa_3 > 0$, and $0 \le \kappa_4 < 1$, $\lambda_{\max}(n^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*) \le \kappa_2$, $E(|\epsilon_i|^{2+\kappa_3}) < \infty$, and $\lim_{n\to\infty}\{\log(v_{q^*})/\log(n)\} = \kappa_4$.

Conditions (C.2) and (C.3) are from Zou and Zhang (2009), which are typically used for establishing the distribution theory of coefficient estimates when $v_{q^*} \to \infty$ as $n \to \infty$. We present the following oracle property.

Theorem 2 (Oracle property). Under Conditions (C.1)–(C.3), if $\phi_n \to \infty$ and $v_{q^*}^2 \phi_n n^{-1/2} \to 0$, then

$$\textit{Consistency in variable selection}: \quad \Pr(\widehat{\mathcal{A}} = \mathcal{A}) \to 1, \tag{6}$$

$$\textit{Asymptotic normality}: \quad \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{1/2}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}\} \xrightarrow{d} \mathrm{Normal}(0,\sigma^2), \tag{7}$$

where $\boldsymbol{\alpha}$ is a vector of norm 1.

Remark 2 (Remark on the orders of p_n and p_0). In this remark, we summarize the orders of p_n and p_0 for the theories in Sections 2.2–2.3. The first one is $\lim_{n\to\infty}\{\log(v_{q^*})/\log(n)\} = \kappa_4$ for $0 \le \kappa_4 < 1$, which is in Condition (C.3). When $p_n = O(n^{1-\zeta})$ for $0 < \zeta \le 1$, the first one is satisfied since $p_n \ge v_{q^*}$. The second one is $v_{q^*}^2 \phi_n n^{-1/2} \to 0$, which, along with $v_{q^*} \ge p_0$, implies $p_0 = o(n^{1/4}\phi_n^{-1/2})$.

2.4. Asymptotic Optimality Under the Scenario That All Candidate Models Are Misspecified

When all candidate models are misspecified, prediction is mainly focused. We will establish the asymptotic optimality of our method in the sense of minimizing the squared prediction loss and risk. Let $\mu_i = E(y_i)$ and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^{\mathrm{T}}$. Different from (1), the data generating process in this subsection is simplified to

$$y_i = \mu_i + \epsilon_i, \quad i = 1,\ldots,n, \tag{8}$$

that is, we do not need μ_i is a linear function of the predictors. The aim of this subsection is to estimate $\boldsymbol{\mu}$ given a predictor matrix $\mathbf{X}_{n\times p_n}$.

For any weight $\mathbf{w} \in \mathcal{W}$, define $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w})$. Write $\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j)^{-1}\mathbf{X}_j^{\mathrm{T}}$ and $\mathbf{P}(\mathbf{w}) = \sum_{j=1}^{q_n} w_j \mathbf{P}_j$. The squared prediction loss and associated risk are defined as

$$L_n(\mathbf{w}) = \|\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$$

and

$$R_n(\mathbf{w}) = E\{L_n(\mathbf{w})\} = \|\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \sigma^2 \mathrm{tr}\{\mathbf{P}^2(\mathbf{w})\}, \tag{9}$$

respectively. Let $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}}\{R_n(\mathbf{w})\}$ and d_n be the column rank of matrix $[\mathbf{X}_1,\ldots,\mathbf{X}_{q_n}]$. The following condition is used to prove asymptotic optimality.

Condition (C.4). $d_n \xi_n^{-1} \to 0$.

Condition (C.4) places restrictions on the rates at which the infimum risk ξ_n and d_n increase with n. A condition that appears to be necessary for Condition (C.4) is that $\xi_n \to \infty$, which is identical to the condition in Hansen (2007) for Mallows weight selection and is quite similar to the conditions of Li (1987) and Andrews (1991); its central role is advocated by Shao (1997). In a typical nested framework, $v_j = j$ and then $d_n = q_n$. In Ando and Li (2014), their conditions (7) and (8) imply that $q_n^2 n^{1/2}\xi_n^{-1} \to 0$, which, along with $d_n = q_n$, further imply our Condition (C.4). In Wan, Zhang, and Zou (2010), instead of Condition (C.4), they assumed that there exists an integer $1 \le G < \infty$ and a positive constant κ such that $E(\epsilon_i^{4G}) \le \kappa$ and $q_n \xi_n^{-2G}\sum_{j=1}^{q_n} R_n(\mathbf{w}_j^o)^G \to 0$, where $\mathbf{w}_j^o$ is a weight vector with the jth element taking on the value of unity and other elements zeros. When $G = 1$ and $d_n = O(q_n)$, their assumption implies that $d_n \xi_n^{-2}\sum_{j=1}^{q_n} R_n(\mathbf{w}_j^o) \to 0$, which along with the fact that $\sum_{j=1}^{q_n} R_n(\mathbf{w}_j^o) \ge \sigma^2 d_n$, leads to Condition (C.4).

Theorem 3 (Asymptotic optimality). If Conditions (C.1) and (C.4) are satisfied, and $\phi_n d_n \xi_n^{-1} \to 0$, then

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}}\{L_n(\mathbf{w})\}} \xrightarrow{p} 1. \tag{10}$$

If, in addition, $\{L_n(\widehat{\mathbf{w}}) - \xi_n\}\xi_n^{-1}$ is uniformly integrable, then

$$\frac{E\{L_n(\widehat{\mathbf{w}})\}}{\inf_{\mathbf{w}\in\mathcal{W}}\{R_n(\mathbf{w})\}} \to 1. \tag{11}$$

Theorem 3 shows that the model averaging procedure is asymptotically optimal in the sense that its squared loss and risk are asymptotically identical to those of the infeasible but best possible model averaging estimator. This property provides a theoretical advantage of our method over selection methods by using AIC or BIC, since the loss or risk of the infeasible but best possible model averaging estimator is smaller or equal to those of the infeasible but best possible model selection estimator.

We note that in proving (10), we actually prove

$$\frac{R_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}}\{R_n(\mathbf{w})\}} \xrightarrow{p} 1, \tag{12}$$

which is also studied for JMA in Hansen and Racine (2012), but is different from (11) because in (12), $\widehat{\mathbf{w}}$ is directly plugged in the right hand of the expression $R_n(\mathbf{w}) = \|\mathbf{P}(\mathbf{w})\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \sigma^2 \mathrm{tr}\{\mathbf{P}^2(\mathbf{w})\}$ (see also (9)). In this article, for the first time, we provide a new type of asymptotic optimally shown by (11), where the randomness of $\widehat{\mathbf{w}}$ is taken into account.

The condition $\phi_n d_n \xi_n^{-1} \to 0$ is very close to Condition (C.4). When $\phi_n = O(1)$, the condition $\phi_n d_n \xi_n^{-1} \to 0$ is implied by Condition (C.4). When $\phi_n \to \infty$, the condition $\phi_n d_n \xi_n^{-1} \to 0$ further restricts the rate of $\phi_n \to \infty$.

Typically, asymptotic optimality is of primary interest when one believes that there does not exist a correctly specified model in candidate model set (see, e.g., Shao 1997; Hansen 2007). Even if a fixed-dimension candidate model is correctly specified, then ξ_n will not increase to ∞ as $n \to \infty$, and thus, the optimality built in this subsection will be invalid. In all, the parsimony and oracle properties built in Sections 2.2–2.3 provide theoretical supports for our model averaging method when there are correct candidate models, while the optimality built here provides a theoretical support for our model averaging method when all candidate models are misspecified.

It is well known that variable selection and optimal risk cannot be achieved simultaneously (Yang 2005), which does not contradict the results in our article because we do not prove the selection consistency shown by Theorem 2 and the optimality shown by Theorem 3 under the same scenarios. The former one is built when there are correct candidate models; while the latter one is built when all candidate models are misspecified.

It is worth pointing out that the optimality or loss efficiency is generally the property of the AIC-type method, whereas our method is the BIC-type method since $\phi_n \to \infty$. The reason why our method still has the optimality mainly relies on the condition $\phi_n d_n \xi_n^{-1} \to 0$, which is similar to $\xi_n \to \infty$, a commonly used condition to derive the optimality for the AIC-type method (Hansen 2007).

Remark 3 (Remark on the order of p_n). In this remark, we summarize the order of p_n for the theories in Section 2.4. By $p_n \geq d_n$, we know that a sufficient condition for the condition $\phi_n d_n \xi_n^{-1} \to 0$ is $p_n = o(\phi_n^{-1}\xi_n)$. Since ξ_n is the infimum of the squared prediction risk $R_n(\mathbf{w})$ and determined by the unknown $\boldsymbol{\mu}$, the order of ξ_n is totally unknown and thus we still keep ξ_n in the sufficient condition of p_n.

2.5. Implementation

A key problem in the implementation of our model averaging method is how to prepare nested candidate models. Claeskens, Croux, and van Kerckhoven (2006) employed a forward selection approach and obtained some nested candidate models. In Ando and Li (2014), the marginal correlation between each predictor and the response variable was used to partition predictors into several groups, and they used these groups to prepare candidate models. Here, we use penalized regression to prepare candidate models. Specifically, we find a solution path of the adaptive LASSO (ALASSO) (Zou 2006), which can be done by the LARS algorithm (Efron et al. 2004). Then we order the predictors depending on the order to enter the solution path in the LARS algorithm, since the order indicates the importance of predictors in some sense. It is worthy to note that a predictor enters the solution path, then may leave and enter the path again, hence we use the last time of entering the path to order this predictor. Last, we set the j candidate model include the first j predictors in the ordering. So we have p_n candidate models.

In the adaptive LASSO, the penalty term is set to be $\lambda \sum_{m=1}^{p_n} \widehat{\beta}_{\mathrm{ols},m}^{-\gamma} |\beta_m|$, that is, we use the least squares estimator $\widehat{\beta}_{\mathrm{ols},m}$ as the initial estimator, which was suggested by Zou (2006). The tuning parameter γ is set as one following much existing literature such as Wang and Leng (2007) and Huang, Ma, and Zhang (2008). From Zou (2006), Zou and Zhang (2009), and Wang, Li, and Leng (2009), when γ is any positive constant, the adaptive LASSO has a selection consistency under some regularity conditions. Varying γ may lead to a different candidate model set by using the solution path and then the smallest correct candidate model, $\inf_{\mathbf{w}\in\mathcal{W}} L_n(\mathbf{w})$ and $\inf_{\mathbf{w}\in\mathcal{W}} R_n(\mathbf{w})$ may change. However, our asymptotic results are all built under given candidate models and the positive tuning parameter γ does not affect the holding of the asymptotic results.

The tuning parameter ϕ_n arises in $\mathcal{S}_n(\mathbf{w})$. When σ^2 is known, the BIC value of the jth candidate model can be simplified to $\|\mathbf{y} - \mathbf{P}_j\mathbf{y}\|^2 + \sigma^2 v_j \log(n)$. Therefore, to keep consistency with BIC, we choose $\phi_n = \log(n)$. When $\mathbf{w} = \mathbf{w}_j^o$, our criterion $\mathcal{S}_n(\mathbf{w})$ reduces to BIC. In addition, when $\phi_n = \log(n)$, the conditions $\phi_n \to \infty$ and $n^{-1/2}\phi_n \to 0$ hold, and Condition (C.4) also holds when d_n and ξ_n have appropriate orders.

When σ^2 is unknown, following Hansen (2007) and Wan, Zhang, and Zou (2010), we use the candidate model with the largest number of predictors to obtain $\widehat{\sigma}^2$, that is, specifically, we define $\widehat{\sigma}^2 = (n - v_{q_n})^{-1}\|\mathbf{y} - \mathbf{X}_{q_n}\widehat{\boldsymbol{\beta}}_{q_n}\|^2$.

Let $\widehat{\boldsymbol{\epsilon}}_j = \mathbf{y} - \mathbf{P}_j\mathbf{y}$ and $\widehat{\mathcal{E}} = (\widehat{\boldsymbol{\epsilon}}_1, \ldots, \widehat{\boldsymbol{\epsilon}}_{q_n})$. Then, $\mathcal{S}_n(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\widehat{\mathcal{E}}^{\mathrm{T}}\widehat{\mathcal{E}}\mathbf{w} + \widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v}\log(n)$. Thus, the minimization of $\mathcal{S}_n(\mathbf{w})$ with respect to $\mathbf{w}$ is a quadratic programming problem. Numerous software packages are available to obtain the solution to this problem (e.g., Matlab and R), and they generally work effectively and efficiently even when q_n is very large.

2.6. Extension to a High-Dimensional Situation

In this section, we extend our method to the situation with a high dimension, that is, $p_n > n$. But we still assume $p_0 < n$, that is, the number of nonzero coefficients is smaller than the sample size. In this scenario, we will combine small candidate models, that is, for any $s \in \{1, \ldots, q_n\}$, $v_s \le n$. This idea is also used in other works on high-dimensional model averaging such as Ando and Li (2014).

First, we focus on the scenario that there are correct candidate models. When the model (1) is correctly specified, this scenario can be easily fulfilled by using a proper model screening method such as the adaptive LASSO combined with BIC which is described in detail at the end of this subsection. It is seen from the proofs that Lemmas 1 and 2 and Theorems 1 and 2 still hold for the high-dimensional situation, because p_n is not used in their proofs and conditions. Even when the overfitted models do not nest each other, they also hold because besides of using (A.2) in the proof of Lemma 1, we can also use that for any $k, j \in \mathcal{O}$ and any constant $c^* > 0$,

$$b_{jk} = a_j - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\mathbf{P}_j\boldsymbol{\epsilon} = a_k - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_j\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\mathbf{P}_j\boldsymbol{\epsilon}$$

and

$$\begin{aligned}
&\Pr\left\{\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\mathbf{P}_j\boldsymbol{\epsilon} + c^*\phi_n(v_{q^*} - v_k) \ge 0\right\} \\
&\quad \le \Pr\left\{\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\boldsymbol{\epsilon} + \|\mathbf{P}_k^{1/2}\boldsymbol{\epsilon}\|^2 + \|\mathbf{P}_k^{1/2}\mathbf{P}_j\boldsymbol{\epsilon}\|^2 \right. \\
&\qquad \left. + c^*\phi_n(v_{q^*} - v_k) \ge 0\right\} \\
&\quad = \Pr\left\{2\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_j\mathbf{P}_k\mathbf{P}_j\boldsymbol{\epsilon} \ge c^*\phi_n(v_k - v_{q^*})\right\} \\
&\quad \le E(2\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_k\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_j\mathbf{P}_k\mathbf{P}_j\boldsymbol{\epsilon})c^{*-1}\phi_n^{-1}(v_k - v_{q^*})^{-1} \\
&\quad = \{2\sigma^2\mathrm{tr}(\mathbf{P}_k) + \sigma^2\mathrm{tr}(\mathbf{P}_k^{1/2}\mathbf{P}_j^2\mathbf{P}_k^{1/2})\}c^{*-1}\phi_n^{-1}(v_k - v_{q^*})^{-1} \\
&\quad \le 3\sigma^2 v_k(v_k - v_{q^*})^{-1}c^{*-1}\phi_n^{-1} \to 0.
\end{aligned}$$

For the asymptotic optimality shown in Theorem 3, since the conditions are all about ξ_n and d_n which are both related to the given candidate models, the asymptotic optimality still holds for the high-dimensional situation.

To implement the proposed method under the high-dimensional situation, we still use the adaptive LASSO to prepare the candidate models. Since $p_n > n$, the least squares estimator is infeasible. We use $\widehat{\boldsymbol{\beta}}_{\mathrm{ini}} = (\widehat{\beta}_{\mathrm{ini},1}, \ldots, \widehat{\beta}_{\mathrm{ini},p_n})^{\mathrm{T}}$ with $\widehat{\beta}_{\mathrm{ini},m} = \sum_{i=1}^n x_{im}y_i/n$ as the initial estimator in the adaptive LASSO. This was proposed by Huang, Ma, and Zhang (2008), from which we know that the adaptive LASSO still has a selection consistency under certain regularity conditions (see their Assumptions A1–A4 and B1–B4 for more details) for the high-dimensional situation with $p_n = O(\exp(n^a))$ for some $0 < a < 1$. To prepare small candidate models for model averaging, we use the first n solutions in the solution path of the adaptive LASSO. Specifically, we first order the predictors selected by BIC depending on the order to enter the solution path, then we use the ordering to prepare nested candidate models. Furthermore, we add the models which belong to the first n solutions and contain the model selected by BIC to the set of candidate models. Finally, we use the candidate model selected by BIC to estimate σ^2, that is, $\widehat{\sigma}^2 = (n - v_{\widehat{j}_{\mathrm{BIC}}})^{-1}\|\mathbf{y} - \mathbf{X}_{\widehat{j}_{\mathrm{BIC}}}\widehat{\boldsymbol{\beta}}_{\widehat{j}_{\mathrm{BIC}}}\|^2$, where $\widehat{j}_{\mathrm{BIC}}$ is the index of the model selected by BIC.

3. Numerical Examples

This section reports the simulation and empirical data results to compare the performance of the PMA with MMA and ALASSO, the latter with the turning parameter chosen by AIC, BIC, or GCV. We applied MMA to exactly the same set of candidate models to which we applied PMA. As suggested by referees, we also compared the proposed method with the smoothly clipped absolute deviation (SCAD) penalty method (Fan and Li 2001) and the minimax concave penalty (MCP) method (Zhang 2010), where the tuning parameters are selected by 5-fold cross-validation.

3.1. Simulation Examples

Example 1 (We have correct candidate models). We generated data from model (1) with $p_n = [2n^{1/2}]$,

$$\boldsymbol{\beta} = \left(\frac{11}{4}\mathbf{1}_{b_n}^{\mathrm{T}}, \mathbf{0}_{b_n}^{\mathrm{T}}, -\frac{13}{9}\mathbf{1}_{b_n}^{\mathrm{T}}, \mathbf{0}_{b_n}^{\mathrm{T}}, -\frac{23}{6}\mathbf{1}_{b_n}^{\mathrm{T}}, \mathbf{0}_{p_n-5b_n}^{\mathrm{T}}\right)^{\mathrm{T}},$$

and $b_n = [n^{1/5}]$. The sample size n was set to vary in $\{100, 200, 400, 600\}$. The predictors $(x_{i1}, \ldots, x_{ip_n})^{\mathrm{T}}, i = 1, \ldots, n$, are iid normal vectors with zero mean and covariance between the mth and kth elements being $0.5^{|m-k|}$. The error term ϵ_i follows Normal$(0, \sigma^2)$ and σ^2 varies such that $R^2 = \mathrm{var}(\sum_{m=1}^{p_n} x_{im}\beta_m)/\mathrm{var}(y_i)$ varies in the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. In this example, the true model can be in the candidate models. Based on 500 replications, we evaluated the methods by comparing the mean squared errors (MSE)

$$\sum_{r=1}^{500}\|\mathbf{X}^{(r)}\widehat{\boldsymbol{\beta}}(\mathbf{w})^{(r)} - \mathbf{X}^{(r)}\widehat{\boldsymbol{\beta}}^{(r)}\|^2/500, \tag{13}$$

where $^{(r)}$ denotes the rth replication. For each parameterization, we normalized the MSEs by dividing the MSE by ALASSO with GCV. The MSEs are shown in Figure 1. We can see that for $n = 100$, the MMA performs the best in the majority of R^2 values, but when $n = 400, 600$, the PMA performs the best in the majority of R^2 values. The MSE pattern by ALASSO with BIC is the closest to that by PMA. For $n = 200, 400, 600$, the PMA outperforms ALASSO with BIC except when n and R^2 are large, which is not unexpected because in this situation, the ALASSO with BIC should have a very high frequency in selecting the true model. When R^2 is large, SCAD and MCP have good performances, but when R^2 is small, their MSEs can be much larger than those of others.

We also reported the mean number of selected variables, shown by Figure 2. It is seen that in most of circumstances, the MMA selects the largest number of variables while the ALASSO with BIC selects the smallest number of variables. The PMA selects slightly more variables than the ALASSO with BIC.

Example 2 (All candidate models are misspecified). This simulation example follows basically the same set-up as in Example 1, except that, here,

$$\begin{aligned}
y_i = &\sin\left(\frac{11}{4}x_{i1}\right) + \cos\left(\frac{11}{4}x_{i2}\right) \\
&+ \sum_{m=3}^{p_n} x_{im}\beta_m + \epsilon_i, \quad i = 1, \ldots, n.
\end{aligned}$$

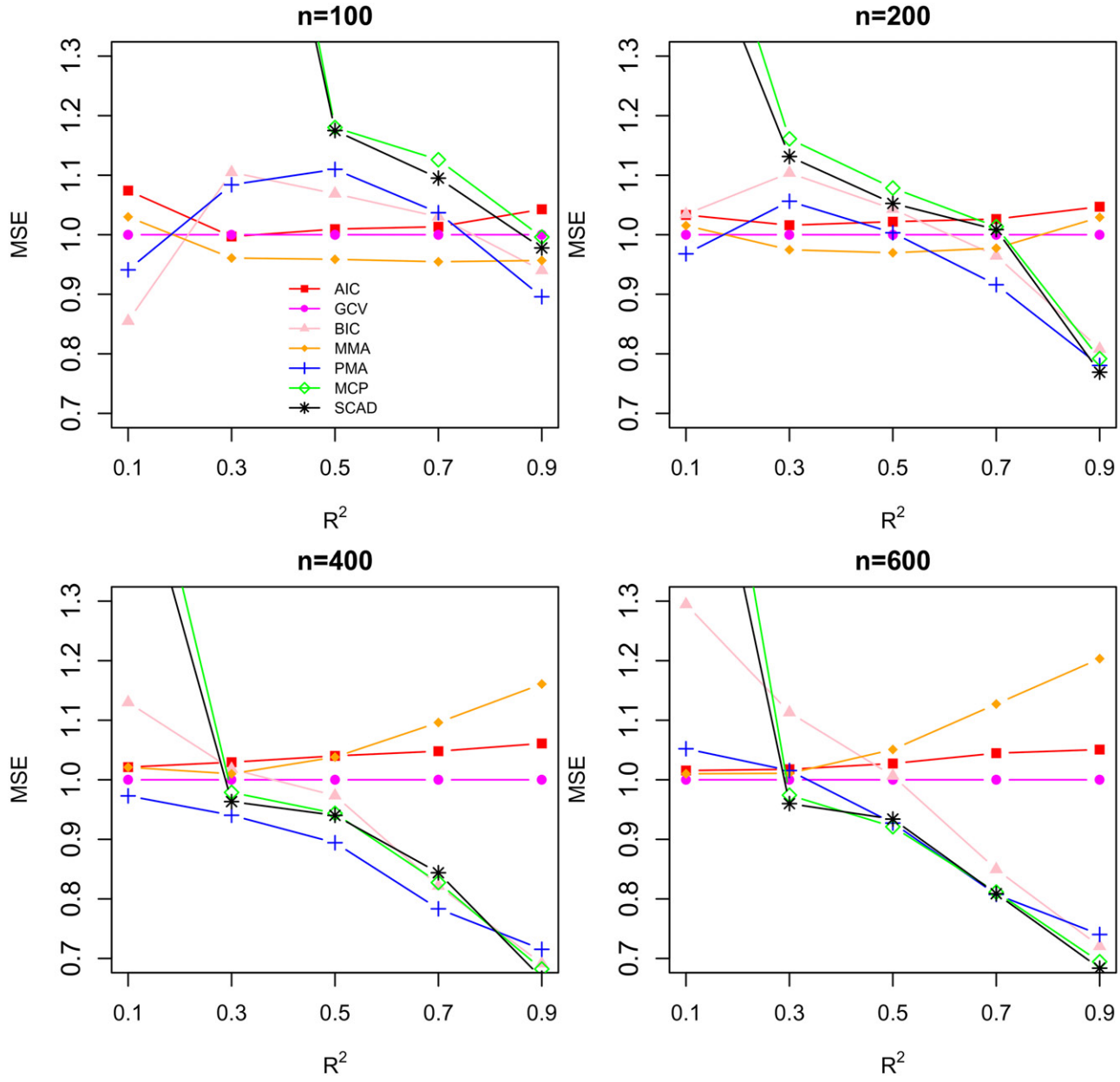

Figure 1. Simulation results for Example 1. MSE against R^2 based on the methods: ALASSO with AIC (filled □), ALASSO with GCV (●), ALASSO with BIC (filled △), MMA(filled ◇), PMA(+), MCP (◇), and SCAD (∗).

In this example, all candidate models are misspecified because the above data generating process has two nonlinear terms. The MSEs are shown in Figure 3. Different from the results in Example 1, for $n = 200, 400, 600$ the PMA outperforms ALASSO with BIC in all R^2 values. The possible reason is that in this example, the data-generating process is not included in the candidate models. Also, the PMA outperforms the MMA, the ALASSO with AIC, and the ALASSO with GCV in all R^2 values except when $n = 100$ and R^2 is moderate, in which the MMA performs the best. Similar to the finding in Example 1, when R^2 is large, SCAD and MCP have good performances but when R^2 is small, their performances are not good.

Example 3. ($p_n > n$) In this example, we set $(n, p_n) \in \{(100, 400), (200, 800)\}$ and $\sigma \in \{1, 2, 4, 8\}$. Other settings are the same as those in Example 1. The MSEs scaled by n and the mean numbers of selected variables are shown in Tables 1 and 2. It is seen that under this high-dimensional setup, the PMA still yields smaller MSEs than other methods in a majority of circumstances. Especially, the MSEs of PMA are much smaller than those of BIC in all circumstances, although in some circumstances, the PMA uses more variables. In all circumstances, the MMA selects the largest number of variables.

3.2. Empirical Example

Example 4. The gene dataset reported by Scheetz et al. (2006) consists of 31,042 genes obtained from 120 rats. The expression levels of gene TRIM32 are the responses. As done by Kim, Choi, and Oh (2008) and Huang, Ma, and Zhang (2008), we excluded genes that were not expressed in the eye or that lacked sufficient

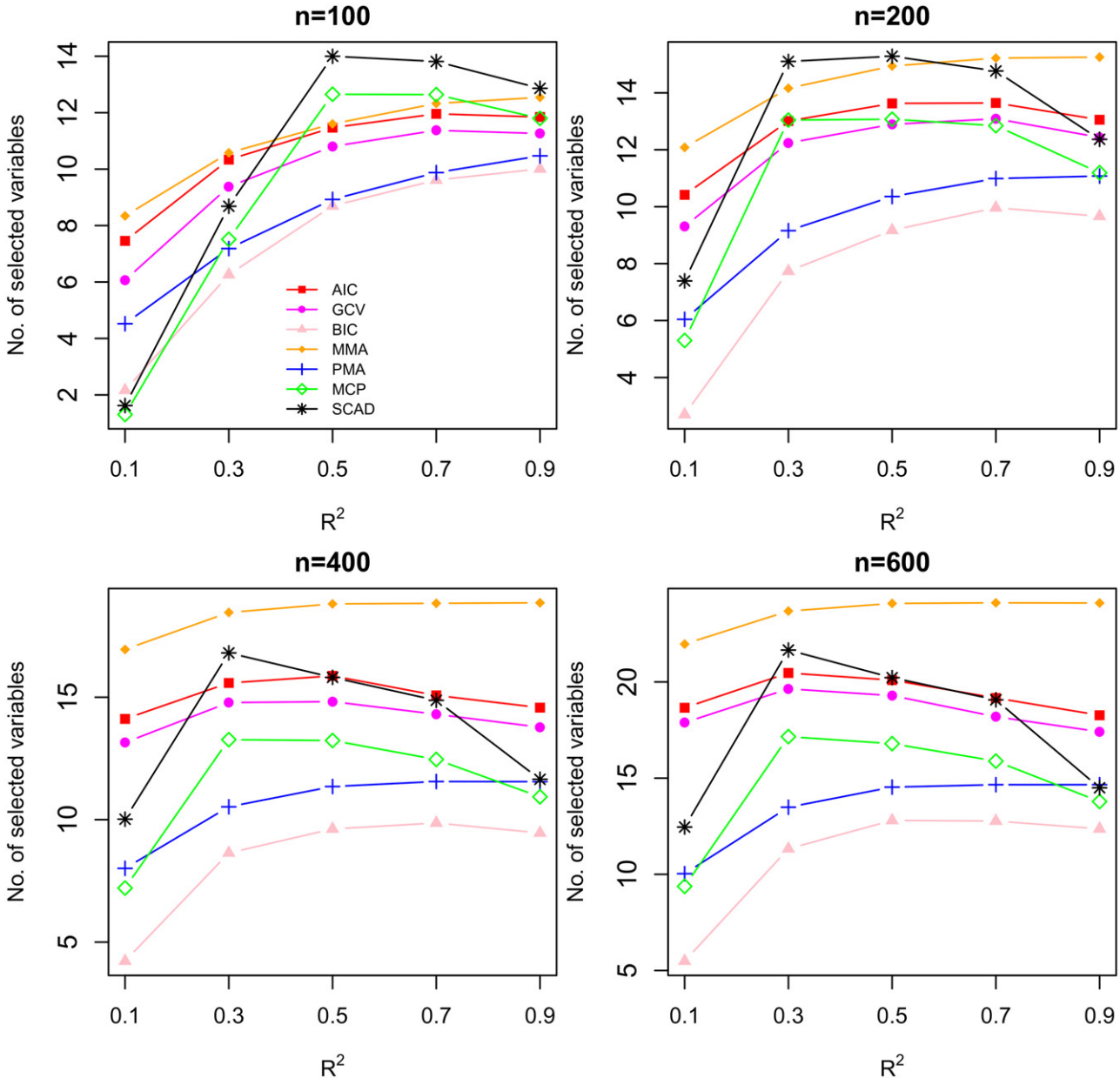

Figure 2. Simulation results for Example 1. The average number of selected variables based on the methods: ALASSO with AIC (filled □), ALASSO with GCV (•), ALASSO with BIC (filled △), MMA(filled ◇), PMA(+), MCP (◇), and SCAD (∗).

variation, standardized all gene expression levels, selected 3000 genes with the largest variance in expression level, and then chose the top p_n genes that have the largest absolute correlation with TRIM32 among the selected 3000 genes. Identically to Huang, Ma, and Zhang (2008), we set p_n to be 100, 200, 300, 400, and 500, and randomly divided the data into a training set of size 80 and a validation set of size 40, doing this 300 times. Note that in this example $p_n > n$, hence we used the procedure in the last paragraph of Section 2.6 to implement the proposed model averaging method.

To compare the results in Huang, Ma, and Zhang (2008), we calculated the median squared errors (MeSE) in predicting responses of validation set and the median numbers of selected variables are shown in Table 3, where we also presented the results from Huang, Ma, and Zhang (2008) in which the tuning parameters are chosen by 5-fold cross-validation. Following that article, their MeSEs are shown with three digits after the decimal point.

The MeSEs by the two model averaging methods are similar, but are smaller than those of the LASSO and ALASSO in all p_n cases. In all p_n cases, the median numbers of selected variables by the PMA are much smaller than those selected by other methods.

4. Discussions

We have proposed a novel model averaging method with a diverging number of parameters. When there are correct models in candidate set, the proposed model averaging estimators of coefficients are parsimonious and have oracle property. When all candidate models are misspecified, the proposed procedure

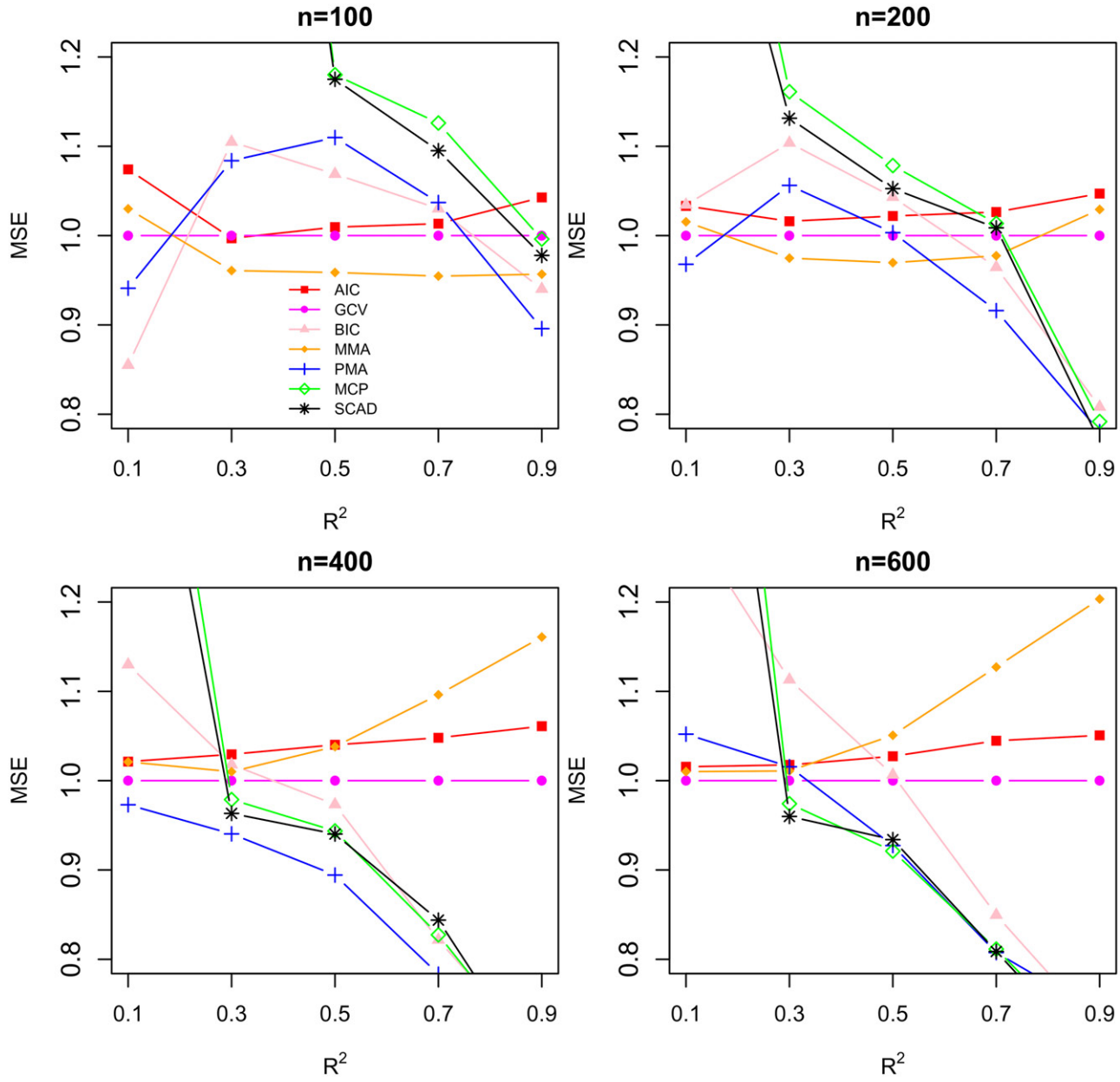

Figure 3. Simulation results for Example 2. MSE against R^2 based on the methods: ALASSO with AIC (filled □), ALASSO with GCV (•), ALASSO with BIC (filled △), MMA(filled ◇), PMA(+), MCP (◇), and SCAD (∗).

Table 1. Simulation results for Example 3: MSE.

(n, p_n)	σ	PMA	MMA	GCV	AIC	BIC	SCAD	MCP
(100,400)	1	0.168	0.214	0.946	0.946	0.960	0.118	0.244
	2	0.755	1.045	1.529	1.480	1.958	0.765	0.917
	4	3.576	4.505	5.558	5.936	7.726	5.241	5.283
	8	16.253	19.521	23.073	34.845	30.148	27.853	26.860
(200,800)	1	0.054	0.131	0.827	0.827	0.828	0.049	0.103
	2	0.330	0.793	1.019	1.014	1.169	0.235	0.349
	4	1.618	3.523	3.197	3.346	4.506	2.279	2.423
	8	8.323	16.628	14.446	25.942	17.812	14.341	15.123

Table 2. Simulation results for Example 3: the average number of selected variables among 300 replications.

(n, p_n)	σ	PMA	MMA	GCV	AIC	BIC	SCAD	MCP
(100,400)	1	10.4	21.8	10.5	10.5	10.2	10.1	10.0
	2	13.5	26.8	19.8	21.4	14.1	18.0	13.8
	4	13.2	28.8	30.3	45.9	13.5	27.4	15.0
	8	12.4	28.9	34.6	69.3	9.0	20.6	10.8
(200,800)	1	10.9	39.5	9.1	9.1	9.1	10.0	10.0
	2	14.5	58.3	15.0	15.2	11.8	13.5	11.7
	4	15.9	59.4	36.9	52.9	12.2	31.5	16.5
	8	17.8	59.1	47.7	114.9	10.4	39.1	18.8

Table 3. MeSE and numbers of selected variables in Example 4 (*gene data*) based on 300 replications.

	MA				Huang, Ma, and Zhang (2008)			
	MMA		PMA		LASSO		ALASSO	
p_n	MeSE	Number	MeSE	Number	MeSE	Number	MeSE	Number
100	0.0043	24	0.0042	13	0.005	20	0.006	18
200	0.0037	21	0.0037	13	0.005	19	0.005	17
300	0.0037	21	0.0037	9	0.005	18	0.005	17
400	0.0036	19	0.0037	9	0.005	22	0.005	19
500	0.0036	20	0.0038	9	0.005	25	0.005	22

NOTES: The "number" column presents the median number of selected variables.

is asymptotically optimal in the sense of minimizing the squared prediction loss. The proposed method and the corresponding theories have been extended to a high-dimensional situation. Numerous studies have also shown its promise.

Our current model averaging method was developed under homoscedastic errors. How to extend it to heteroscedastic case warrants further research. In addition, extending our method and theory to a general nonnested framework is a very important topic of future research. Last, how to extend our method to generalized linear models remains for a future research. The recent work of Ando and Li (2017) may serve as a useful guide in this regard.

Appendix

A.1. Proof of Lemma 1

Based on definitions at the beginning of Section 2.2, we know that the first $q^* - 1$ models are underfitted and the last $q_n - q^*$ models are overfitted. Let $\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j^\top \mathbf{X}_j)\mathbf{X}_j^\top a_j = \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{y}$, $b_{jk} = \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{I}_n - \mathbf{P}_k)\mathbf{y}$ and $\mathbf{\Phi}$ be an $q_n \times q_n$ matrix with the jkth element $\Phi_{jk} = b_{jk} + \phi_n \widehat{\sigma}^2(v_j + v_k)/2$. By simple calculations, we have $\mathcal{S}_n(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}$. It is straightforward to show that

$$b_{kj} = b_{jk} = a_{\max\{k,j\}}. \tag{A.1}$$

Using the fact that $\phi_n \to \infty$ and Markov's inequality, we observe that, for any $k \in \mathcal{O}$ and any constant $c^* > 0$,

$$\begin{aligned}
&\Pr\left\{a_{q^*} - a_k + c^*\phi_n(v_{q^*} - v_k) \geq 0\right\} \\
&= \Pr\left\{\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}_k - \mathbf{P}_{q^*})\boldsymbol{\epsilon} + c^*\phi_n(v_{q^*} - v_k) \geq 0\right\} \\
&= \Pr\left\{\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}_k - \mathbf{P}_{q^*})\boldsymbol{\epsilon} \geq c^*\phi_n(v_k - v_{q^*})\right\} \\
&\leq E\{\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}_k - \mathbf{P}_{q^*})\boldsymbol{\epsilon}\}c^{*-1}\phi_n^{-1}(v_k - v_{q^*})^{-1} \\
&= \sigma^2 c^{*-1}\phi_n^{-1} \to 0.
\end{aligned} \tag{A.2}$$

Let $\widetilde{\mathbf{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_{q_n})^{\mathrm{T}}$ be a weight vector belonging to $\mathcal{W}$ with $\delta = \sum_{j=q^*+1}^{q_n} \widetilde{w}_j > 0$. Partition $\widetilde{\mathbf{w}}$ as $\widetilde{\mathbf{w}} = (\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}, \widetilde{w}_{q^*}, \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}})^{\mathrm{T}}$, according to which we partition $\mathbf{\Phi}$ as

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_{11} & \mathbf{\Phi}_{12} & \mathbf{\Phi}_{13} \\ \mathbf{\Phi}_{21} & \mathbf{\Phi}_{22} & \mathbf{\Phi}_{23} \\ \mathbf{\Phi}_{31} & \mathbf{\Phi}_{32} & \mathbf{\Phi}_{33} \end{pmatrix}, \tag{A.3}$$

where $\mathbf{\Phi}_{22} = \Phi_{q^*q^*}$. Let

$$\overline{\mathbf{w}} = (\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}, \widetilde{w}_{q^*} + \delta, \mathbf{0}_{(q_n - q^*)\times 1}^{\mathrm{T}})^{\mathrm{T}} \quad \text{and} \quad \mathbf{t} = (\mathbf{0}_{(q^*-1)\times 1}^{\mathrm{T}}, -\delta, \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}})^{\mathrm{T}}.$$

Then, $\overline{\mathbf{w}} = \widetilde{\mathbf{w}} - \mathbf{t}$. A direct simplification yields that

$$\begin{aligned}
\mathcal{S}_n(\overline{\mathbf{w}}) &= \overline{\mathbf{w}}^{\mathrm{T}}\mathbf{\Phi}\overline{\mathbf{w}} = (\widetilde{\mathbf{w}} - \mathbf{t})^{\mathrm{T}}\mathbf{\Phi}(\widetilde{\mathbf{w}} - \mathbf{t}) = \mathcal{S}_n(\widetilde{\mathbf{w}}) + (\mathbf{t} - 2\widetilde{\mathbf{w}})^{\mathrm{T}}\mathbf{\Phi}\mathbf{t} \\
&= \mathcal{S}_n(\widetilde{\mathbf{w}}) + \{(\mathbf{0}_{(q^*-1)\times 1}^{\mathrm{T}}, -\delta, \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}})^{\mathrm{T}} - 2\widetilde{\mathbf{w}}\}^{\mathrm{T}} \\
&\quad \mathbf{\Phi}(\mathbf{0}_{(q^*-1)\times 1}^{\mathrm{T}}, -\delta, \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}})^{\mathrm{T}} \\
&= \mathcal{S}_n(\widetilde{\mathbf{w}}) + (-2\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}, -\delta - 2\widetilde{w}_{q^*}, -\widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}) \\
&\quad \mathbf{\Phi}(\mathbf{0}_{(q^*-1)\times 1}^{\mathrm{T}}, -\delta, \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}})^{\mathrm{T}} \\
&= \mathcal{S}_n(\widetilde{\mathbf{w}}) + 2\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{12}\delta - 2\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{13}\widetilde{\mathbf{w}}_{(3)} + (\delta^2 + 2\widetilde{w}_{q^*}\delta)\mathbf{\Phi}_{22} \\
&\quad -(\delta + 2\widetilde{w}_{q^*})\mathbf{\Phi}_{23}\widetilde{\mathbf{w}}_{(3)} + \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}\mathbf{\Phi}_{32}\delta - \widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}\mathbf{\Phi}_{33}\widetilde{\mathbf{w}}_{(3)} \\
&= \mathcal{S}_n(\widetilde{\mathbf{w}}) - 2(\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{13}\widetilde{\mathbf{w}}_{(3)} - \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{12}\delta) \\
&\quad -2\widetilde{w}_{q^*}(\mathbf{\Phi}_{23}\widetilde{\mathbf{w}}_{(3)} - \delta\mathbf{\Phi}_{22}) \\
&\quad -(\widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}\mathbf{\Phi}_{33}\widetilde{\mathbf{w}}_{(3)} - \delta^2\mathbf{\Phi}_{22}).
\end{aligned} \tag{A.4}$$

It is seen from (A.1) that

$$\begin{aligned}
&\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{13}\widetilde{\mathbf{w}}_{(3)} - \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{12}\delta = \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}(\mathbf{\Phi}_{13}\widetilde{\mathbf{w}}_{(3)} - \mathbf{\Phi}_{12}\delta) \\
&= \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\left(\sum_{j=q^*+1}^{q_n} \Phi_{1j}\widetilde{w}_j - \Phi_{1q^*}\delta, \ldots,\right. \\
&\qquad \left.\sum_{j=q^*+1}^{q_n} \Phi_{(q^*-1)j}\widetilde{w}_j - \Phi_{(q^*-1)q^*}\delta\right)^{\mathrm{T}} \\
&= \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\left[\sum_{j=q^*+1}^{q_n} \{a_j + \phi_n\widehat{\sigma}^2(v_1 + v_j)/2\}\widetilde{w}_j\right. \\
&\quad \left. -\{a_{q^*} + \phi_n\widehat{\sigma}^2(v_1 + v_{q^*})/2\}\delta,\right.
\end{aligned}$$

$$
\begin{aligned}
&\ldots, \sum_{j=q^*+1}^{q_n} \{a_j + \phi_n\widehat{\sigma}^2(v_{q^*-1} + v_j)/2\}\widetilde{w}_j \\
&-\{a_{q^*} + \phi_n\widehat{\sigma}^2(v_{q^*-1} + v_{q^*})/2\}\delta\Big]^{\mathrm{T}} \\
&= \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\Bigg(\sum_{j=q^*+1}^{q_n} [\{a_j + \phi_n\widehat{\sigma}^2(v_1 + v_j)/2 \\
&-\{a_{q^*} + \phi_n\widehat{\sigma}^2(v_1 + v_{q^*})/2\}]\widetilde{w}_j, \\
&\ldots, \sum_{j=q^*+1}^{q_n} [\{a_j + \phi_n\widehat{\sigma}^2(v_{q^*-1} + v_j)/2\} \\
&-\{a_{q^*} + \phi_n\widehat{\sigma}^2(v_{q^*-1} + v_{q^*})/2\}]\widetilde{w}_j\Bigg)^{\mathrm{T}} \\
&= \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\Bigg[\sum_{j=q^*+1}^{q_n} \{a_j - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j - v_{q^*})/2\}\widetilde{w}_j, \\
&\ldots, \sum_{j=q^*+1}^{q_n} \{a_j - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j - v_{q^*})/2\}\widetilde{w}_j\Bigg]^{\mathrm{T}},
\end{aligned}
$$

which together with (A.2) and the first part of Condition (C.1), implies

$$
\Pr\{\widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{13}\widetilde{\mathbf{w}}_{(3)} - \widetilde{\mathbf{w}}_{(1)}^{\mathrm{T}}\mathbf{\Phi}_{12}\delta \geq 0\} \to 1. \tag{A.5}
$$

Similarly, from (A.1), we have

$$
\begin{aligned}
\mathbf{\Phi}_{23}\widetilde{\mathbf{w}}_{(3)} - \delta\mathbf{\Phi}_{22} &= \sum_{j=q^*+1}^{q_n} \Phi_{q^*j}\widetilde{w}_j - \delta\Phi_{q^*q^*} \\
&= \sum_{j=q^*+1}^{q_n} (\Phi_{q^*j} - \Phi_{q^*q^*})\widetilde{w}_j \\
&= \sum_{j=q^*+1}^{q_n} \{a_j - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j - v_{q^*})/2\}\widetilde{w}_j,
\end{aligned}
$$

which together with (A.2) and the first part of Condition (C.1), implies

$$
\Pr\{\mathbf{\Phi}_{23}\widetilde{\mathbf{w}}_{(3)} - \delta\mathbf{\Phi}_{22} \geq 0\} \to 1. \tag{A.6}
$$

From (A.1), we have

$$
\begin{aligned}
&\widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}\mathbf{\Phi}_{33}\widetilde{\mathbf{w}}_{(3)} - \delta^2\mathbf{\Phi}_{22} \\
&= \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\Phi_{jk} - (\sum_{j=q^*+1}^{q_n} \widetilde{w}_j)^2\Phi_{q^*q^*} \\
&= \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k(\Phi_{jk} - \Phi_{q^*q^*}) \\
&= \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\{a_{\max\{j,k\}} - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j + v_k - 2v_{q^*})/2\} \\
&= \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\phi_n\widehat{\sigma}^2(v_k - v_{q^*})/4 \\
&\quad + \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\phi_n\widehat{\sigma}^2(v_j - v_{q^*})/4 \\
&\quad + \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\{a_{\max\{j,k\}} - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j + v_k - 2v_{q^*})/4\} \\
&\geq \phi_n\widehat{\sigma}^2/4 \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k + \phi_n\widehat{\sigma}^2/4 \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k \\
&\quad + \sum_{j=q^*+1}^{q_n}\sum_{k=q^*+1}^{q_n} \widetilde{w}_j\widetilde{w}_k\{a_{\max\{j,k\}} - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j + v_k - 2v_{q^*})/4\} \\
&= \delta^2\phi_n\widehat{\sigma}^2/2 + \sum_{j=q^*+1}^{q_n}\sum_{j\leq k\leq q_n} \widetilde{w}_j\widetilde{w}_k \\
&\quad \times\{a_k - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j + v_k - 2v_{q^*})/4\} \\
&\quad + \sum_{j=q^*+1}^{q_n}\sum_{q^*+1\leq k<j} \widetilde{w}_j\widetilde{w}_k\{a_j - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j + v_k - 2v_{q^*})/4\} \\
&\geq \delta^2\phi_n\widehat{\sigma}^2/2 + \sum_{j=q^*+1}^{q_n}\sum_{j\leq k\leq q_n} \widetilde{w}_j\widetilde{w}_k\{a_k - a_{q^*} + \phi_n\widehat{\sigma}^2(v_k - v_{q^*})/4\} \\
&\quad + \sum_{j=q^*+1}^{q_n}\sum_{q^*+1\leq k<j} \widetilde{w}_j\widetilde{w}_k\{a_j - a_{q^*} + \phi_n\widehat{\sigma}^2(v_j - v_{q^*})/4\},
\end{aligned}
$$

which together with (A.2), $\delta > 0$, and the first part of Condition (C.1), implies

$$
\Pr\{\widetilde{\mathbf{w}}_{(3)}^{\mathrm{T}}\mathbf{\Phi}_{33}\widetilde{\mathbf{w}}_{(3)} - \delta^2\mathbf{\Phi}_{22} > 0\} \to 1. \tag{A.7}
$$

Using (A.4)–(A.7), we know that when $\delta > 0$, $\Pr\{\mathcal{S}(\overline{\mathbf{w}}) < \mathcal{S}(\widetilde{\mathbf{w}})\} \to 1$, which and the fact that $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}} \mathcal{S}_n(\mathbf{w})$ lead to

$$
\Pr\Bigg(\sum_{j=q^*+1}^{q_n} \widehat{w}_j = 0\Bigg) \to 1. \tag{A.8}
$$

By combining (A.8) and the fact that $\widehat{w}_k \geq 0$, we obtain (3).

A.2. Proof of Lemma 2

It is seen that

$$
\begin{aligned}
&E\{(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})\} \\
&= \sigma^2\mathrm{tr}\{\mathbf{X}^*(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\mathrm{T}}\} = v_{q^*}\sigma^2,
\end{aligned}
$$

which together with $v_{q^*}\phi_n n^{-1} \to 0$, implies

$$
n^{-1}(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta}) = o_p(1). \tag{A.9}
$$

From (A.9) and Condition (C.2), a direct calculation shows that, for any $s \in \mathcal{U}$,

$$
\begin{aligned}
&n^{-1}(a_s - a_{q^*}) \\
&= n^{-1}(\mathbf{y}^{\mathrm{T}}\mathbf{P}_{q^*}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{P}_s\mathbf{y}) \\
&= n^{-1}(\widehat{\boldsymbol{\beta}}_{q^*}^{\mathrm{T}}\mathbf{\Pi}_{q^*}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \widehat{\boldsymbol{\beta}}_s^{\mathrm{T}}\mathbf{\Pi}_s\mathbf{X}_s^{\mathrm{T}}\mathbf{X}_s\mathbf{\Pi}_s^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s) \\
&= n^{-1}(\widehat{\boldsymbol{\beta}}_{q^*}^{\mathrm{T}}\mathbf{\Pi}_{q^*}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \widehat{\boldsymbol{\beta}}_s^{\mathrm{T}}\mathbf{\Pi}_{q^*}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s) \\
&= n^{-1}(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}) \\
&= n^{-1}\{\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{q^*} - \boldsymbol{\beta})\}^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^* \\
&\quad \times\{\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{q^*} - \boldsymbol{\beta})\} \\
&\geq 2^{-1}n^{-1}(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta}) \\
&\quad -n^{-1}(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*(\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*} - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta}) \\
&\geq 2^{-1}\lambda_{\min}(n^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)\|\mathbf{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_s - \mathbf{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta}\|^2
\end{aligned}
$$

$$
\begin{aligned}
&-n^{-1}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})\\
&\geq 2^{-1}\lambda_{\min}(n^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})\min_{j\in\mathcal{A}}(\beta_j^2)\\
&\quad -n^{-1}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})\\
&\geq 2^{-1}\kappa_1\min_{j\in\mathcal{A}}(\beta_j^2)\\
&\quad -n^{-1}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}(\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{q^*}-\boldsymbol{\Pi}_{q^*}^{\mathrm{T}}\boldsymbol{\beta})\\
&= 2^{-1}\kappa_1\min_{j\in\mathcal{A}}(\beta_j^2)+o_p(1). \qquad \text{(A.10)}
\end{aligned}
$$

By (A.10) and $v_{q^*}\phi_n n^{-1}\to 0$, we obtain that, for any positive constant c^* and $s\in\mathcal{U}$,

$$
\Pr\{a_{q^*}-a_s+\phi_n c^*(v_{q^*}-v_s)\geq 0\}\to 0. \qquad \text{(A.11)}
$$

Let $k\in\{1,\ldots,q^*-1\}$ and $\overline{\mathbf{w}}=(\widehat{w}_1,\ldots,\widehat{w}_{k-1},0,\widehat{w}_{k+1},\ldots,\widehat{w}_{q^*}+\widehat{w}_k,\widehat{w}_{q^*+1},\ldots\widehat{w}_{q_n})^{\mathrm{T}}$. Then,

$$
\begin{aligned}
\mathcal{S}_n(\overline{\mathbf{w}}) &= \overline{\mathbf{w}}^{\mathrm{T}}\boldsymbol{\Phi}\overline{\mathbf{w}}=\{\widehat{\mathbf{w}}^{\mathrm{T}}+(0,\ldots,0,-\widehat{w}_k,0,\ldots,0,\widehat{w}_k,0,\ldots,0)\}\boldsymbol{\Phi}\\
&\quad\times\{\widehat{\mathbf{w}}+(0,\ldots,0,-\widehat{w}_k,0,\ldots,0,\widehat{w}_k,0,\ldots,0)^{\mathrm{T}}\}\\
&= \mathcal{S}_n(\widehat{\mathbf{w}})+\widehat{w}_k^2(\Phi_{q^*q^*}+\Phi_{kk}-\Phi_{q^*k}-\Phi_{kq^*})\\
&\quad+2\widehat{w}_k\textstyle\sum_{j=1}^{q_n}\{\widehat{w}_j(\Phi_{q^*j}-\Phi_{kj})\}\\
&= \mathcal{S}_n(\widehat{\mathbf{w}})+\widehat{w}_k^2(\Phi_{q^*q^*}+\Phi_{kk}-\Phi_{q^*k}-\Phi_{kq^*})\\
&\quad+2\widehat{w}_k^2(\Phi_{q^*k}-\Phi_{kk})\\
&\quad+2\widehat{w}_k w_{q^*}(\Phi_{q^*q^*}-\Phi_{kq^*})+2\widehat{w}_k\sum_{j>q^*}\{\widehat{w}_j(\Phi_{q^*j}-\Phi_{kj})\}\\
&\quad+2\widehat{w}_k\sum_{q^*>j>k}\{\widehat{w}_j(\Phi_{q^*j}-\Phi_{kj})\}\\
&\quad+2\widehat{w}_k\sum_{1\leq j<k}\{\widehat{w}_j(\Phi_{q^*j}-\Phi_{kj})\}. \qquad \text{(A.12)}
\end{aligned}
$$

When $j>q^*$, from (3), (A.1), and Condition (C.1) we have

$$
\Pr\{\widehat{w}_j(\Phi_{q^*j}-\Phi_{kj})=0\}=\Pr\{\widehat{w}_j\phi_n\widehat{\sigma}^2(v_{p^*}-v_k)/2=0\}\to 1.
$$

When $k<j<q^*$, from the second part of Condition (C.1), (A.1), and (A.11) we have

$$
\Pr\{\Phi_{q^*j}-\Phi_{kj}\geq 0\}=\Pr\{a_{q^*}-a_j+\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)/2\geq 0\}\to 0.
$$

Similarly, when $1\leq j<k$, we have

$$
\Pr\{\Phi_{q^*j}-\Phi_{kj}\geq 0\}=\Pr\{a_{q^*}-a_k+\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)/2\geq 0\}\to 0.
$$

In addition, from (A.1), we have

$$
\begin{aligned}
&2\widehat{w}_k\widehat{w}_{q^*}(\Phi_{q^*q^*}-\Phi_{kq^*})+\widehat{w}_k^2(\Phi_{q^*q^*}+\Phi_{kk}-\Phi_{q^*k}-\Phi_{kq^*})\\
&\quad+2\widehat{w}_k^2(\Phi_{q^*k}-\Phi_{kk})\\
&\quad=\widehat{w}_k\widehat{w}_{q^*}\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)+\widehat{w}_k^2(a_k-a_{q^*})\\
&\quad+2\widehat{w}_k^2\{a_{q^*}-a_k+\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)/2\}\\
&\quad=(\widehat{w}_k\widehat{w}_{q^*}+\widehat{w}_k^2)\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)-\widehat{w}_k^2(a_k-a_{q^*}).
\end{aligned}
$$

From above results and $\mathcal{S}_n(\widehat{\mathbf{w}})\leq\mathcal{S}_n(\overline{\mathbf{w}})$, we have

$$
\Pr\{(\widehat{w}_k\widehat{w}_{q^*}+\widehat{w}_k^2)\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)-\widehat{w}_k^2(a_k-a_{q^*})\geq 0\}\to 1,
$$

and thus when $\widehat{w}_k\neq 0$, $\widehat{w}_k\leq(a_k-a_{q^*})^{-1}(\widehat{w}_{q^*}+\widehat{w}_k)\phi_n\widehat{\sigma}^2(v_{q^*}-v_k)$ holds with probability approaching to 1, which together with (A.10), implies (5).

A.3. Proof of Theorem 2

From Theorem 3.3 of Zou and Zhang (2009) and Conditions (C.2) and (C.3), we have that when $j=q^*$,

$$
\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},j}-\boldsymbol{\beta}_{\mathcal{A}})\xrightarrow{d}\text{Normal}(0,\sigma^2), \qquad \text{(A.13)}
$$

and when $j\in\mathcal{U}$,

$$
\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},j}-\boldsymbol{\beta}_{\mathcal{A}})=O_p(n^{1/2}). \qquad \text{(A.14)}
$$

From $v_{q^*}^2\phi_n n^{-1/2}\to 0$, (5) and (A.14), we have

$$
\sum_{j\in\mathcal{U}}w_j\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A},j}-\boldsymbol{\beta}_{\mathcal{A}})=o_p(1),
$$

which, along with (4) and (A.13), implies (7) and thus

$$
\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{1/2}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}\}=O_p(1). \qquad \text{(A.15)}
$$

From (4) and (A.15), we can obtain (6). From (A.13) and (6), we obtain (7).

A.4. Proof of Theorem 3

Write $\mathbf{H}=(\mathbf{X}_1,\ldots,\mathbf{X}_{q_n})$, $\mathbf{P}_{\mathbf{H}}=\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-}\mathbf{H}^{\mathrm{T}}$, and

$$
\mathcal{S}_n^*(\mathbf{w})=\mathcal{S}_n(\mathbf{w})-\|\boldsymbol{\epsilon}\|^2-2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\epsilon}+2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}.
$$

Since $\|\boldsymbol{\epsilon}\|^2+2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\epsilon}-2\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}$ is unrelated to $\mathbf{w}$, we have $\widehat{\mathbf{w}}=\text{argmin}_{\mathbf{w}\in\mathcal{W}}\,\mathcal{S}_n^*(\mathbf{w})$. By simple calculations, we obtain that

$$
\mathcal{S}_n^*(\mathbf{w})-L_n(\mathbf{w})=\phi_n\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v}-2\left\{\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}+\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}-\boldsymbol{\mu}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}\right\}
$$

and

$$
R_n(\mathbf{w})-L_n(\mathbf{w})=\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2-\sigma^2\text{tr}\{\mathbf{P}^2(\mathbf{w})\}-2\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{I}_n-\mathbf{P}(\mathbf{w})\}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}.
$$

So, as in the proof of Theorem 1′ in Wan, Zhang, and Zou (2010), in order to prove (10), we need only to verify that

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}}\{R_n^{-1}(\mathbf{w})(\phi_n\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v})\}\xrightarrow{p}0, && \text{(A.16a)}\\
&\sup_{\mathbf{w}\in\mathcal{W}}\{R_n^{-1}(\mathbf{w})\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\}\xrightarrow{p}0, && \text{(A.16b)}\\
&\sup_{\mathbf{w}\in\mathcal{W}}[R_n^{-1}(\mathbf{w})|\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{P}(\mathbf{w})-\mathbf{P}_{\mathbf{H}}\}\boldsymbol{\epsilon}|]\xrightarrow{p}0, && \text{(A.16c)}\\
&\sup_{\mathbf{w}\in\mathcal{W}}\{R_n^{-1}(\mathbf{w})\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2\}\xrightarrow{p}0, && \text{(A.16d)}\\
&\sup_{\mathbf{w}\in\mathcal{W}}[R_n^{-1}(\mathbf{w})\text{tr}\{\mathbf{P}^2(\mathbf{w})\}]\to 0, && \text{(A.16e)}
\end{aligned}
$$

and

$$
\sup_{\mathbf{w}\in\mathcal{W}}[R_n^{-1}(\mathbf{w})|\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{I}_n-\mathbf{P}(\mathbf{w})\}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}|]\xrightarrow{p}0. \qquad \text{(A.16f)}
$$

First, by Condition (C.1) and $\phi_n d_n\xi_n^{-1}\to 0$, we have (A.16a). Using the fact that the $\mathbf{P}_j$'s are symmetric and idempotent, we have $\text{tr}[\mathbf{P}^2(\mathbf{w})]\leq\text{tr}[\mathbf{P}(\mathbf{w})]\leq d_n$, so (A.16e) is implied by Condition (C.4). By the Markov inequality, for any $\delta>0$,

$$
\begin{aligned}
\Pr(\xi_n^{-1}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}>\delta)&\leq E(\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon})\xi_n^{-1}\delta^{-1}\\
&=\sigma^2 d_n\xi_n^{-1}\delta^{-1}\to 0. \qquad \text{(A.17)}
\end{aligned}
$$

Since $\mathbf{P}_j$'s and $\mathbf{P}_{\mathbf{H}}-\mathbf{P}_j$'s are symmetric and idempotent, we have

$$
\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2\leq\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\leq\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon},
$$

so (A.16b) and (A.16d) are implied by Condition (C.4) and (A.17). From (9), it is straightforward to show that

$$
R_n^{-2}(\mathbf{w})[\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{I}_n-\mathbf{P}(\mathbf{w})\}\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}]^2
$$

$$\le R_n^{-2}(\mathbf{w})\|\{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\boldsymbol{\mu}\|^2\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2$$
$$\le R_n^{-1}(\mathbf{w})\|\mathbf{P}(\mathbf{w})\boldsymbol{\epsilon}\|^2;$$

from this and (A.16d), we see that (A.16f) is true. Lastly, for (A.16c), it follows from (9) that

$$R_n^{-2}(\mathbf{w})\left[\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{P}(\mathbf{w}) - \mathbf{P}_{\mathbf{H}}\}\boldsymbol{\epsilon}\right]^2 = R_n^{-2}(\mathbf{w})\left[\boldsymbol{\mu}^{\mathrm{T}}\{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}\right]^2$$
$$\le R_n^{-2}(\mathbf{w})\|\{\mathbf{I}_n - \mathbf{P}(\mathbf{w})\}\boldsymbol{\mu}\|^2\|\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}\|^2$$
$$\le R_n^{-1}(\mathbf{w})\|\mathbf{P}_{\mathbf{H}}\boldsymbol{\epsilon}\|^2; \quad \text{(A.18)}$$

from this and (A.17), we obtain (A.16c). This completes the proof of (10).

Next, we prove (11). Using (A.16d)–(A.16f), we see that

$$\sup_{\mathbf{w}\in\mathcal{W}}\{|L_n(\mathbf{w}) - R_n(\mathbf{w})|R_n^{-1}(\mathbf{w})\} \xrightarrow{p} 0. \quad \text{(A.19)}$$

It is straightforward to show that

$$L_n(\widehat{\mathbf{w}})\xi_n^{-1} = \sup_{\mathbf{w}\in\mathcal{W}}[\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\}\{L_n(\mathbf{w})R_n^{-1}(\mathbf{w})\}]. \quad \text{(A.20)}$$

By (10), (A.19), and (A.20), we have

$$L_n(\widehat{\mathbf{w}})\xi_n^{-1} \le \sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\}\sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\mathbf{w})R_n^{-1}(\mathbf{w})\}$$
$$\le \sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\} \times[1 + \sup_{\mathbf{w}\in\mathcal{W}}\{|L_n(\mathbf{w}) - R_n(\mathbf{w})|R_n^{-1}(\mathbf{w})\}]$$
$$= 1 + o_p(1)$$

and

$$L_n(\widehat{\mathbf{w}})\xi_n^{-1} \ge \sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\}\inf_{\mathbf{w}\in\mathcal{W}}\{L_n(\mathbf{w})R_n^{-1}(\mathbf{w})\}$$
$$= \sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\} \times\left(1 + \inf_{\mathbf{w}\in\mathcal{W}}[\{L_n(\mathbf{w}) - R_n(\mathbf{w})\}R_n^{-1}(\mathbf{w})]\right)$$
$$\ge \sup_{\mathbf{w}\in\mathcal{W}}\{L_n(\widehat{\mathbf{w}})L_n^{-1}(\mathbf{w})\} \times\left[1 - \sup_{\mathbf{w}\in\mathcal{W}}\{|L_n(\mathbf{w}) - R_n(\mathbf{w})|R_n^{-1}(\mathbf{w})\}\right]$$
$$= 1 + o_p(1).$$

Therefore, $L_n(\widehat{\mathbf{w}})\xi_n^{-1} \xrightarrow{p} 1$, from which we have $\{L_n(\widehat{\mathbf{w}}) - \xi_n\}\xi_n^{-1} = o_p(1)$. Consequently, from the uniform integrability of $\{L_n(\widehat{\mathbf{w}}) - \xi_n\}\xi_n^{-1}$, it is obvious that $E[\{L_n(\widehat{\mathbf{w}}) - \xi_n\}\xi_n^{-1}] \to 0$, by which we obtain (11).

Acknowledgments

The authors thank the co-editor, an associate editor, and two referees for their insightful suggestions and comments that have substantially improved an earlier version of this article.

Funding

Zhang was supported by the National Natural Science Foundation of China (NNSFC) (Grant nos. 71522004, 11471324, and 71631008). Zou was supported by NNSFC (Grant no. 11331011) and the Ministry of Science and Technology of China (Grant no. 2016YFB0502301). Liang was supported by NSF grant DMS-1620898, and Award Number 11529101 made by NNSFC. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

References

Ando, T., and Li, K.-C. (2014), "A Model-Averaging Approach for High-Dimensional Regression," *Journal of the American Statistical Association*, 109, 254–265. [972,974,975,976]

——— (2017), "A Weight-Relaxed Model Averaging Approach for High-Dimensional Generalized Linear Models," *The Annals of Statistics*, 45, 2654–2679. [980]

Andrews, D. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [974]

Buckland, S., Burnham, K., and Augustin, N. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [972]

Claeskens, G., Croux, C., and van Kerckhoven, J. (2006), "Variable Selection for Logistic Regression Using a Prediction-Focused Information Criterion," *Biometrics*, 62, 972–979. [975]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [975]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [976]

Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [972,973,974,975]

——— (2014), "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530. [973]

Hansen, B. E., and Racine, J. (2012), "Jacknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [972,975]

Hjort, N., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [972]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [972]

Huang, J., Ma, S. G., and Zhang, C. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618. [975,976,977,978,980]

Ishwaran, H., and Rao, J. S. (2003), "Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455. [973]

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly Clipped Absolute Deviation on High Dimensions," *Journal of the American Statistical Association*, 103, 1665–1673. [977]

Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [972]

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_l, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [974]

Liang, H., Zou, G. H., Wan, A. T. K., and Zhang, X. Y. (2011), "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106, 1053–1066. [972]

Liu, C.-A. (2015), "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159. [972]

Liu, Q., and Okui, R. (2013), "Heteroskedasticity-Robust C_p Model Averaging," *Econometrics Journal*, 16, 462–473. [972,973]

Lu, X., and Su, L. (2015), "Jackknife Model Averaging for Quantile Regressions," *Journal of Econometrics*, 188, 40–58. [972]

Miller, A. J. (2002), *Subset Selection in Regression* (2nd ed.), London: Chapman and Hall. [972]

Raftery, A. E., and Zheng, Y. (2003), "Discussion: Performance of Bayesian Model Averaging," *Journal of the American Statistical Association*, 98, 931–938. [973]

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006), "Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease," *Proceedings of the National Academy of Sciences of the United States of America*, 103, 14429–14434. [977]

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264. [974,975]

Wan, A. T. K., Zhang, X., and Zou, G. H. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [973,974,975,982]

Wang, H., and Leng, C. (2007), "Unified Lasso Estimation via Least Squares Approximation," *Journal of the American Statistical Association*, 101, 1418–1429. [975]

Wang, H. S., Li, B., and Leng, C. L. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters," *Journal of the Royal Statistical Society*, Series B, 71, 671–683. [975]

Yang, Y. H. (2005), "Can the Strengths of AIC and BIC Be Shared?—A Conflict Between Model Identification and Regression Estimation," *Biometrika*, 92, 937–950. [975]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214. [972,973]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [976]

Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [972]

Zhang, X., Wan, A., and Zhou, Z. (2012), "Focused Information Criteria, Model Selection and Model Averaging in a Tobit Model With a Non-Zero Threshold," *Journal of Business and Economic Statistics*, 30, 132–142. [972]

Zhang, X., Zou, G., and Liang, H. (2014), "Model Averaging and Weight Choice in Linear Mixed-Effects Models," *Biometrika*, 101, 205–218. [972]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [975]

Zou, H., and Zhang, H. H. (2009), "On the Adaptive Elastic-Net With a Diverging Number of Parameters," *The Annals of Statistics*, 37, 1733–1751. [974,975,982]