JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 00, NO. 0, 1–10
https://doi.org/10.1080/07350015.2022.2116442

A Scalable Frequentist Model Averaging Method

Rong Zhu[a], Haiying Wang[b], Xinyu Zhang[c], and Hua Liang[d]

[a]Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai, China; [b]Department of Statistics, University of Connecticut, Storrs, CT; [c]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; [d]Department of Statistics, George Washington University, Washington, DC

ABSTRACT

Frequentist model averaging is an effective technique to handle model uncertainty. However, calculation of the weights for averaging is extremely difficult, if not impossible, even when the dimension of the predictor vector, p, is moderate, because we may have 2^p candidate models. The exponential size of the candidate model set makes it difficult to estimate all candidate models, and brings additional numeric errors when calculating the weights. This article proposes a scalable frequentist model averaging method, which is statistically and computationally efficient, to overcome this problem by transforming the original model using the singular value decomposition. The method enables us to find the optimal weights by considering at most p candidate models. We prove that the minimum loss of the scalable model averaging estimator is asymptotically equal to that of the traditional model averaging estimator. We apply the Mallows and Jackknife criteria to the scalable model averaging estimator and prove that they are asymptotically optimal estimators. We further extend the method to the high-dimensional case (i.e., $p \geq n$). Numerical studies illustrate the superiority of the proposed method in terms of both statistical efficiency and computational cost.

ARTICLE HISTORY
Received February 2022
Accepted August 2022

KEYWORDS
Asymptotic optimality; High-dimensional data; Jackknife criterion; Mallows criterion; Singular value decomposition

1. Introduction

Model averaging is a popular and effective approach in dealing with model uncertainty and improving prediction accuracy. Instead of picking a single "best" model according to some model assessment criterion in traditional model selection, the model averaging approach advocates the pooling of predictions by giving higher weights to better models. The approach often reduces the risk in regression analysis, as "betting" on multiple models prevent the case of a singly selected model being poor (Leung and Barron 2006).

In the existing literature, Bayesian model averaging has been well studied, and there is a large amount of literature on this approach; see Hoeting et al. (1999), Raftery et al. (1997), and the references therein. As an alternative, frequentist model averaging on which this article focuses has been gaining increasing attention. Buckland et al. (1997) suggested using weights based on exponential Akaike information criterion (AIC) (Akaike 1973) to combine estimates from different candidate models. Yang (2001) and Yuan and Yang (2005) proposed a mixing estimator. Hjort and Claeskens (2003) provided an asymptotic analysis of model average estimators in the likelihood-based framework. Hansen (2007) proposed selecting the optimal weights for model averaging by minimizing a Mallows criterion. Liang et al. (2011) developed a procedure for selecting optimal weights such that the resultant estimator has the minimum mean-squared error. Hansen and Racine (2012) proposed a jackknife model averaging method using leave-one-out cross-validation.

When there are p predictors, there are 2^p candidate models to consider. Thus, the size of the candidate model set is usually enormous, which causes critical issues for obtaining optimal weights by minimizing a criterion that requires to estimate all candidate models. For example, given 20 predictors, there are more than one million candidate models. It would be computationally difficult to estimate one million models and to obtain the optimal weights for averaging over them. Calculating so many weights may also cause the loss of prediction efficiency due to computational error. Attempts have been made to reduce this burden in the literature. One approach is to consider a subset of all possible models by assuming a nested structure among candidate models (Hansen 2007; Hansen and Racine 2012). However, this nested model structure is not applicable in some practical problems. For example, in labor economics, it is not suitable to assume that possible predictors can be expressed in a nested way (Wooldridge 2003).

In this article, we develop a scalable frequentist model averaging method based on singular value decomposition (SVD), from which, we obtain the left singular vectors of the predictor matrix, fit regression models on these singular vectors separately to get the estimators, and then average these estimators to obtain a model averaging estimator. This strategy enables us to find optimal weights by considering at most p candidate models

CONTACT Hua Liang hliang@gwu.edu Department of Statistics, George Washington University, Washington, DC.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

instead of 2^p candidate models, and greatly advocates the applicability of frequentist model averaging. Magnus and Durbin (1999) proposed the weighted-average least squares estimator which transformed covariates based on the relation between the least squares estimators of interested parameters and that of nuisance parameters. The idea of orthogonalization was also used in Clyde et al. (1996) for Bayesian model mixing. They developed a Bayesian framework for orthogonalized design matrix assuming that the normal linear regression model is a correct model. Similar strategy can also be found in Charkhi et al. (2016), Jolliffe (1982), and Park (1981). We adopt a pure frequentist approach and we do not assume that the full model is correct. Compared with the traditional model averaging based on the original covariates, the improvement of prediction efficiency from the scalable model averaging, as well as computational gain from orthogonalization, are remarkable.

We prove that the minimum loss of the scalable model averaging estimator is asymptotically equal to that of the traditional model averaging estimator. This indicates that the best performance of the scalable model averaging estimator is asymptotically as good as the best performance of the traditional model averaging estimator.

Furthermore, we use Mallows (Hansen 2007) and Jackknife (Hansen and Racine 2012) weight selections methods to illustrate the proposed scalable model averaging, and prove that the resulting scalable Mallows/Jackknife model averaging estimators are asymptotically optimal, an oracle property established in the literature (Hansen 2007; Hansen and Racine 2012; Ando and Li 2014).

Another contribution of our scalable model averaging is its applicability to high-dimensional data ($p \geq n$, with n being the sample size). Claeskens (2012) suggested averaging estimators from penalization-based estimation approaches, but she did not investigate how to average these estimators. Ando and Li (2014) developed a model averaging approach for high-dimensional regression. They took an average over candidate models that are formed by grouping the covariates according to marginal correlations. In this article we also apply our strategy to the high-dimensional case, and provide three practical approaches. The first procedure is to reduce the dimensionality by screening the original covariates; the second procedure is to reduce the number of left singular vectors by removing the ones with small singular values; and the third procedure is to reduce the number of left singular vectors by the sure independent ranking and screening (SIRS) method (Zhu 2011). We thank an anonymous reviewer for suggesting the third idea. A comparison with Ando and Li's (2014) approach indicates that our method achieves a higher estimation efficiency with remarkably lower computational cost.

The reminder of the article is organized as follows. Section 2 proposes our scalable model averaging method, and show the asymptotic equivalence in the minimum loss to the traditional model averaging. Section 3 uses the Mallows and Jackknife criteria to illustrate the scalable model averaging and establishes the asymptotic optimality of the resulting estimators. Section 4 extends the scalable model averaging method to the high-dimensional case. Sections 5 and 6 present numerical evidences from simulated and real datasets. Proofs of the Theorems and additional numerical results are presented in the supplementary materials.

2. Scalable Model Averaging

2.1. Model Averaging Estimators

Let $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ be a random sample, where y_i is the response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the predictor vector. Our working model is the linear regression model:

$$y_i = \mu_i + e_i = \mathbf{x}_i^\top \boldsymbol{\theta} + e_i = \sum_{j=1}^{p} \theta_j x_{ij} + e_i,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is the regression coefficient vector and e_i's are uncorrelated and heteroscedastic model errors such that $E(e_i|\mathbf{x}_i) = 0$ and $E(e_i^2|\mathbf{x}_i) = \sigma_i^2$. In matrix notation,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $\mathbf{e} = (e_1, \ldots, e_n)^\top$, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is the $n \times p$ predictor matrix. In this and next two sections, we focus on the case of $n \gg p$.

When we are not sure which of the p predictor variables should be included in a model, there are up to 2^p candidate models to consider. The method of model averaging is to average all estimates from each individual candidate model. Let M be the number of candidate models to be considered for averaging. Since 2^p may be too large to handle computationally, all 2^p possible candidate models are not always included for averaging, for example, the nested model structure (Hansen 2007; Hansen and Racine 2012) only considered a small proportion of all the possible candidate models. Thus, in traditional model averaging methods, it is typically assumed that $M \ll 2^p$. If all possible candidate models are included for averaging, then $M = 2^p$. For $m \leq M$, let $\mathbf{X}_m$ be the predictors matrix corresponding to the mth candidate model. The ordinary least square (OLS) estimator of $\boldsymbol{\theta}$ is $\tilde{\boldsymbol{\theta}}_m = (\mathbf{X}_m^\top \mathbf{X}_m)^{-1} \mathbf{X}_m^\top \mathbf{y}$, and the estimator of $\boldsymbol{\mu}$ from the mth candidate is $\tilde{\boldsymbol{\mu}}_m = \mathbf{X}_m \tilde{\boldsymbol{\theta}}_m$. Let $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_M)^\top$ be the weight vector in the unit simplex in $\mathbb{R}^M$:

$$\tilde{\mathcal{H}}^M = \left\{ \tilde{w}_i \in [0, 1] : \sum_{m=1}^{M} \tilde{w}_m = 1 \right\}. \tag{2}$$

A model averaging estimator of $\boldsymbol{\mu}$ is

$$\tilde{\boldsymbol{\mu}}(\tilde{\mathbf{w}}) = \sum_{m=1}^{M} \tilde{w}_m \tilde{\boldsymbol{\mu}}_m. \tag{3}$$

To choose the weight vector $\tilde{\mathbf{w}}$, Mallows model averaging (Hansen 2007) and Jackknife model averaging (Hansen and Racine 2012) are commonly used and are proved to be efficient. Regardless of Mallows model averaging and Jackknife model averaging, the target function to optimize is a quadratic function of $\tilde{\mathbf{w}}$, for which the numerical solutions have been thoroughly studied and algorithms are widely available.

2.2. Scalable Model Averaging Estimators

Now we introduce the scalable model averaging method to address the computational challenge caused by the exponential size of the candidate model set. The idea is to use SVD to convert the predictor matrix into a column-orthogonal predictor matrix.

Denote the SVD of matrix $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, such that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$, where $\mathbf{U}$ is an $n \times p$ column-orthogonal matrix, $\mathbf{D}$ is a $p \times p$ rectangular diagonal matrix with nonnegative real numbers on the diagonal, $\mathbf{V}$ is a $p \times p$ orthogonal matrix, and $\mathbf{I}_p$ is the p-dimensional identity matrix. The original model (1) can be represented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e} = \mathbf{U}\boldsymbol{\beta} + \mathbf{e}, \tag{4}$$

where $\boldsymbol{\beta} = \mathbf{D}\mathbf{V}^\top\boldsymbol{\theta}$.

Let u_{ij} be ijth element of $\mathbf{U}$. All the information in the original predictors x_{xj}'s for the responses is preserved in u_{ij}'s. We can see this from the angle of mean prediction through the hat matrix, which is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ for the original predictors. This matrix converts the observed responses to the estimated mean responses. The hat matrix from $\mathbf{U}$ is identical to $\mathbf{H}$ because

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top. \tag{5}$$

Thus, the prediction from the transformed model (4) is equivalent to that from the original model (1).

The model (4) can be rewritten as

$$y_i = \sum_{j=1}^{p} \beta_j u_{ij} + e_i, \quad i = 1, \ldots, n. \tag{6}$$

Notice there are still 2^p candidate models in the transformed model (6). Let $\mathbf{U}_m$ be the predictor matrix corresponding to the mth candidate model in the transformed model (6), and denote the predictor set of the mth candidate model as

$$S_m = \{j : 1 \le j \le p,\ j \text{ is included in the } m\text{th candidate model}\}.$$

The OLS estimator $\hat{\boldsymbol{\beta}}_m$ of the regression coefficient from any candidate model with $\mathbf{U}_m$ consists of the OLS estimators from univariate regressions of $\mathbf{y}$ on each column of $\mathbf{U}_m$. To see this explicitly, let $\hat{\beta}_j$ be the OLS estimator of β_j from the model with a single predictor $\mathbf{u}_{(j)}$, the jth column of $\mathbf{U}$. We have

$$\hat{\beta}_j = \left\{\sum_{i=1}^{n} u_{ij}^2\right\}^{-1} \sum_{i=1}^{n} u_{ij} y_i = \sum_{i=1}^{n} u_{ij} y_i = \mathbf{u}_{(j)}^\top \mathbf{y}.$$

On the other hand, the OLS estimator $\hat{\boldsymbol{\beta}}_m$ from the mth model with $\mathbf{U}_m$ is

$$\hat{\boldsymbol{\beta}}_m = (\mathbf{U}_m^\top\mathbf{U}_m)^{-1}\mathbf{U}_m^\top\mathbf{y} = \mathbf{U}_m^\top\mathbf{y} = (\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_{p_m}})^\top, \tag{7}$$

where $j_1, \ldots, j_{p_m}$ are elements of the set S_m, and p_m is the size of S_m. Since columns of $\mathbf{U}$ are orthogonal, the marginal estimator for each β_j remains the same as the jth component of the full model estimator. Note that the transform between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in (4) is one-on-one when $\mathbf{X}$ is column full rank, so we can transform regression coefficient estimates based on $\mathbf{U}_m$ to that based on the original predictors to interpret the results. From (7), the OLS estimator of any candidate model consists of $\{\hat{\beta}_j : j \in S_m\}$. Thus, the prediction based on the mth candidate model $\mathbf{U}_m$ is

$$\hat{\mu}_m = \mathbf{U}_m\hat{\boldsymbol{\beta}}_m = \sum_{j \in S_m} \mathbf{u}_{(j)}\hat{\beta}_j.$$

For a model averaging estimator, it can be written as a linear combination of these $\mathbf{u}_{(j)}\hat{\beta}_j$'s, namely,

$$\begin{aligned}\hat{\boldsymbol{\mu}}(\tilde{\mathbf{w}}) = \sum_{m=1}^{M} \tilde{w}_m \hat{\mu}_m &= \sum_{m=1}^{M} \tilde{w}_m \sum_{j \in S_m} \mathbf{u}_{(j)}\hat{\beta}_j \\ &= \sum_{j=1}^{p} \left\{\sum_{m=1}^{M} \tilde{w}_m I_{S_m}(j)\right\} \mathbf{u}_{(j)}\hat{\beta}_j, \end{aligned} \tag{8}$$

where $I_{S_m}(j)$ is the indicator function of the set S_m, that is, $I_{S_m}(j) = 1$ if $j \in S_m$ and 0 otherwise. From Equation (8), if we define

$$w_j = \sum_{m=1}^{M} \tilde{w}_m I_{S_m}(j) \in [0, 1] \tag{9}$$

to be the weight, we can obtain the model averaging estimators under the framework of model (4) by combining (averaging) the predictions from the p univariate regression models. Let $\mathbf{w} = (w_1, \ldots, w_p)^\top$ be the weight vector in the simplex

$$\mathcal{H} = \left\{w_j \in [0, 1] : j = 1, \ldots, p\right\}.$$

The constraint $\sum_{j=1}^{p} w_j = 1$ is not needed in the converted problem because $\sum_{j=1}^{p} w_j = \sum_{m=1}^{M}\sum_{j=1}^{p} \tilde{w}_m I_{S_m}(j)$ is typically not 1. Since the estimator of $\boldsymbol{\mu}$ in the jth univariate regression model is $\hat{\boldsymbol{\mu}}_{(j)} = \mathbf{u}_{(j)}\hat{\beta}_j$, the scalable model averaging estimate of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{j=1}^{p} w_j \hat{\boldsymbol{\mu}}_{(j)} = \sum_{j=1}^{p} w_j \mathbf{u}_{(j)}\hat{\beta}_j. \tag{10}$$

A direct benefit is that we just need to choose the weights of size p rather than those of size 2^p in traditional model averaging. Since the left singular vector matrix has the same prediction ability as the original predictor matrix as shown in (5), our scalable model averaging may keep the efficiency of the traditional model averaging using much less computational cost. Meanwhile, reducing the size of the weights from 2^p to p may improve the efficiency of frequentist model averaging for practical application because the reduced weight size may reduce the numerical error in optimization. These are observed in numerical studies with synthetic and real datasets.

It is worthwhile to mention that an element of $\tilde{\mathbf{w}}$ is the weight of a candidate model in the original predictors, but an element of $\mathbf{w}$ does not have this interpretation. Unlike the case of the regression coefficient, the linear transformation from $\tilde{\mathbf{w}}$ to $\mathbf{w}$ in (9) is not invertible; we cannot obtain $\tilde{\mathbf{w}}$ from $\mathbf{w}$. Thus, we lose the intuitive interpretability of the estimated weight on the original candidate models. However, such a loss is not a concern when the focus is on the accuracy of prediction or estimator.

2.3. Asymptotical Equivalence of the Minimum Loss

Intuitively, one may expect to pay a price in terms of predictive efficiency for the huge computational gain. Interestingly, our method does not pay this price in the large sample sense. We prove that the minimum loss of our scalable model averaging estimator is asymptotically equal to that of the model averaging estimator based on the 2^p candidate models with the original covariates. In contrast, existing model averaging methods often assume that the number of candidate models is much smaller than 2^p for concerning computational burden, but it is unclear whether this causes any inflation in the minimum loss. The minimum loss of model averaging estimators in a reduced candidate model space is in general larger unless excluded models are truly redundant.

Let $\tilde{L}(\tilde{\mathbf{w}}) = \|\tilde{\boldsymbol{\mu}}(\tilde{\mathbf{w}}) - \boldsymbol{\mu}\|^2$, where $\tilde{\mathbf{w}} \in \tilde{\mathcal{H}}$, be the sum of squared loss of the model-averaging estimator $\tilde{\boldsymbol{\mu}}$ in (3) using the original covariates, and $L(\mathbf{w}) = \|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$, where $\mathbf{w} \in \mathcal{H}$, be the sum of squared loss of a scalable model averaging estimator $\hat{\boldsymbol{\mu}}(\mathbf{w})$ in (10). Denote $\xi_n = \inf_{\mathbf{w}\in\mathcal{H}} E\{L(\mathbf{w})\}$. We require the following conditions. For some constants c_1, c_2, and c_3,

Condition C.1. $E(e_i^4) \le c_1 < \infty$, for $i = 1, \ldots, n$.

Condition C.2. $p/\xi_n \to 0$ as $n \to \infty$,

Remark. C.1 is a quite mild condition which requires that the fourth moments of errors are bounded. C.2 imposes that the dimension should be a lower order term with respect to the infimum of the risk of $\hat{\boldsymbol{\mu}}(\mathbf{w})$. This assumption is reasonable in our framework. The condition indicates that no linear model consisting of available covariate is a correct model, that is, all candidate models are wrong. When no correct model is included, it often holds that $L(\mathbf{w}) = O_P(n)$, and C.2 holds. A similar assumption for a moderate dimension case was imposed and discussed in Hansen (2007) and Hansen and Racine (2012).

Theorem 1. Under Conditions C.1 and C.2,

$$\frac{\inf_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w})}{\inf_{\tilde{\mathbf{w}}\in\tilde{\mathcal{H}}^{2^p}} \tilde{L}(\tilde{\mathbf{w}})} \to 1, \tag{11}$$

in probability, as $n \to \infty$, where $\tilde{\mathcal{H}}^{2^p}$ is the weight set defined in (2) with $M = 2^p$.

Theorem 1 shows that the minimum loss for the scalable model averaging estimator in (10) is asymptotically the same as that for the traditional model averaging estimator in (3) using the original covariates with 2^p candidate models. It provides a theoretical foundation on asymptotical comparison of the scalable model averaging estimator and the traditional model averaging estimator. For the traditional model averaging estimator, if $M = 2^p$, that is, all possible candidate models are included for averaging, we need to estimate 2^p candidate models and calculate 2^p weights, while for our scalable model averaging estimator we just need to fit p univariate regressions and calculate a weight vector of size p instead of 2^p. The only extra cost is to perform a SVD on the predictor matrix $\mathbf{X}$. Thus, our method greatly reduces the computational cost. Furthermore, this can be done without paying any penalty in terms of the estimation efficiency. This result is intuitive: on the one hand, the left singular vectors of the SVD on the design matrix are linear combinations of predictors and there is no information loss in the transformation as shown in (5); on the other hand, the orthogonality of the singular vectors greatly reduces the weight size from 2^p to p.

3. Scalable MMA and JMA

Theorem 1 in the previous section shows that the scalable model averaging estimator and the traditional model averaging estimator achieve the same minimum loss asymptotically. We often obtain the weight vector $\mathbf{w}$ in practice by minimizing some criterion $\mathcal{C}(\mathbf{w})$, that is

$$\hat{\mathbf{w}}_{\text{MA}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathcal{C}(\mathbf{w}). \tag{12}$$

We consider two specific choices of $\mathcal{C}(\mathbf{w})$ to illustrate our scalable estimator: the Mallows model averaging (MMA) (Hansen 2007) and the Jackknife Model Averaging (JMA) (Hansen and Racine 2012). We summarize the scalable model averaging approach in Algorithm 1.

The MMA is based on the Mallows criterion. Since the number of covariates used in univariate regression model is 1, the Mallows criterion of our scalable model averaging is

$$\mathcal{M}(\mathbf{w}) = \|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \mathbf{y}\|^2 + 2\hat{\sigma}^2\mathbf{w}^\top\mathbf{1}, \tag{13}$$

where $\|\cdot\|^2$ stands for the Euclidean norm and $\hat{\sigma}^2 = (n-p)^{-1}\|\mathbf{y} - \mathbf{U}\mathbf{U}^\top\mathbf{y}\|^2$. From (13), the weight vector is obtained as $\hat{\mathbf{w}}_{\text{MMA}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathcal{M}(\mathbf{w})$.

The JMA is based on the leave-one-out cross validation criterion. Let $\mathbf{H}_j = \mathbf{u}_{(j)}\mathbf{u}_{(j)}^\top$, and $h_{j,ii}$ be the ith diagonal element of $\mathbf{H}_j$, that is, $h_{j,ii} = u_{ij}^2$. Define $\mathbf{D}_j$ to be the diagonal matrix with $(1-h_{j,ii})^{-1}$ being its ith diagonal element, and let $\mathbf{H}_j^{\text{JMA}} = \mathbf{D}_j(\mathbf{H}_j - \mathbf{I}_n) + \mathbf{I}_n$ and $\mathbf{H}^{\text{JMA}}(\mathbf{w}) = \sum_{j=1}^p w_j\mathbf{H}_j^{\text{JMA}}$. Following Hansen and Racine (2012), the Jackknife criterion is

$$\mathcal{J}(\mathbf{w}) = \|\mathbf{H}^{\text{JMA}}(\mathbf{w})\mathbf{y} - \mathbf{y}\|^2. \tag{14}$$

From (14), the weight vector is obtained as $\hat{\mathbf{w}}_{\text{MMA}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \mathcal{M}(\mathbf{w})$.

Algorithm 1 Scalable model averaging

1. Obtain $\mathbf{U}$, the matrix of left-singular vectors of $\mathbf{X}$, by SVD decomposition: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
2. For each column of $\mathbf{U}$, $\mathbf{u}_{(j)}$, calculate the least square prediction $\hat{\boldsymbol{\mu}}_j = \mathbf{u}_{(j)}\hat{\beta}_j$
3. Calculate the scalable model averaging estimator

$$\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}}) = \sum_{j=1}^{p} \hat{w}_j\hat{\boldsymbol{\mu}}_{(j)}$$

by averaging the predictions $\hat{\boldsymbol{\mu}}_j$, $j = 1, \ldots, p$, where $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_p)^\top$ is the weight vector obtained from data. Two specific choices are $\hat{\mathbf{w}}_{\text{MMA}}$ or $\hat{\mathbf{w}}_{\text{JMA}}$ obtained from Mallows' model averaging or Jackknife model averaging.

Our scalable Mallows/Jackknife model averaging estimators are optimal in the sense that they asymptotically achieve the

minimum loss of the infeasible best possible model averaging estimator using *the original model (1)*. This can be seen by combining Theorem 1 (establishing the equivalence of the minimum losses from the original and scalable model averaging) and existing results that MMA and JMA achieve the minimum loss for a given set of candidate models (Hansen 2007; Hansen and Racine 2012; Ando and Li 2014). Nevertheless, the orthonormality of the left singular vectors can be used to simplify the proof and weaken the required assumptions specified below.

Condition C.3. $\max_{i,j} u_{ij}^2 \le c_2 n^{-1}$ for some constant c_2.

Condition C.4. $\|\boldsymbol{\mu}\|^2/n \le c_3 < \infty$ for some constant c_3.

Remark. C.3 means that no individual element of u_{ij}'s dominates all the others. Since $\sum_{i=1}^n u_{ij}^2 = 1$ for all j, this assumption is also reasonable. C.4 is quite mild since the equation of $\|\boldsymbol{\mu}\|^2 = O(n)$ often holds.

Theorem 2. Under Conditions C.1–C.4,

$$\frac{L(\hat{\mathbf{w}}_{\text{MMA}})}{\inf_{\tilde{\mathbf{w}}\in\tilde{\mathcal{H}}^{2^p}} \tilde{L}(\tilde{\mathbf{w}})} \to 1 \quad \text{and} \quad \frac{L(\hat{\mathbf{w}}_{\text{JMA}})}{\inf_{\tilde{\mathbf{w}}\in\tilde{\mathcal{H}}^{2^p}} \tilde{L}(\tilde{\mathbf{w}})} \to 1,$$

in probability, as $n \to \infty$.

4. Extension to High-Dimensional Data

In this section, we extend the application of the scalable model averaging method to the high-dimensional setting with $p \ge n$. In order to extend the scalable model averaging method to high-dimensional data, we propose three practical procedures to address the curse of dimensionality. The first procedure is to reduce the dimension by screening the original predictors using the sure independent ranking and screening (SIRS) method (Zhu et al. 2011). The second procedure is to reduce the number of left singular vectors by removing the ones corresponding to small singular values. The third procedure is to perform the SIRS on left singular vectors. For our first procedure with high-dimensional data, we prefer the SIRS over the sure independence screening (SIS) proposed by Fan and Lv (2008) because we assume that all candidate models are wrong, which is the case for most practical problems. The SIRS allows us to assume that no linear candidate model is correct, and we just need to assume that the distribution of the response depends only on some of the covariates (active covariate) and does not depend on other covariates (inactive covariates). Following Zhu et al. (2011), the SIRS screens the covariates based on the magnitude of the following statistics instead of the marginal correlation, $\tilde{\omega}_j = \frac{1}{n}\sum_{i=1}^n \left\{\frac{1}{n}\sum_{l=1}^n x_{ij} I_{(-\infty, y_i)}(y_l)\right\}^2$. Derivation and interpretation of this statistics can be found in Zhu et al. (2011). This procedure is similar to the SIS used in the first step of Ando and Li (2014). Theorems 2 and 3 in Zhu et al. (2011) indicate that with probability approaching one, the SIRS can reduce the dimensionality without losing any active covariate. Thus, using SIRS preprocessing allows us to keep all candidate models that involve any activate covariates in a reduced dimension. Once the dimension is reduced, the asymptotic results for the low-dimensional case in Section 2 hold.

The second procedure for high-dimensional data is to drop some left singular vectors with small singular values such that we keep k left singular vectors with largest singular values. We do so by following the factor model setting where the columns of $\mathbf{U}$ with small singular values are assumed to be independent on the response (Bai 2003). For real high-dimensional data, it is almost always that $\text{rank}(\mathbf{X}) = n$, but it is also often true that many nonzero singular values are small or even close to 0. Hence, it is possible, though probably hard, to establish the asymptotic equivalence of the minimum loss for model averaging with the original covariates under the factor model setting, but it is out of the scope of this article and is a future investigation topic. Empirically, we show that this procedure works well for finite sample sizes in Sections 5 and 6.

In the last procedure for high-dimensional data, we screen left singular vectors by the SIRS and drop some left singular vectors that are least relevant to $\mathbf{y}$. Again, it is difficult to establish asymptotic equivalence of the minimum loss, because we would have to guarantee that the loss due to SIRS on $\mathbf{U}$ is asymptotically zero. Empirically, we will show that this procedure works well in Sections 5 and 6.

We extend Algorithm 1 to account for high-dimensional data, and summarize it as Algorithm 2.

Algorithm 2 A scalable model averaging for high-dimensional data

Step 1 (a) Choose k covariates by SIRS to get the subset $\mathbf{X}_k$ from $\mathbf{X}$, and then obtain the left singular vectors from the SVD of $\mathbf{X}_k$;
or (b) obtain $\mathbf{U}$ from a SVD and keep the its k columns with the largest k singular values;
or (c) obtain $\mathbf{U}$ from a SVD and choose k columns of it using the SIRS method.

Step 2 For each left singular vector from Step 1, say $\mathbf{u}_{(j)}$, calculate the least square prediction $\hat{\boldsymbol{\mu}}_j = \mathbf{u}_{(j)}\hat{\beta}_j$. Calculate the scalable model averaging estimator

$$\hat{\boldsymbol{\mu}}(\hat{\mathbf{w}}) = \sum_{j=1}^{k} \hat{w}_j \hat{\boldsymbol{\mu}}_{(j)}$$

by averaging the predictions $\hat{\boldsymbol{\mu}}_j$, $j = 1, \ldots, k$, where $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_p)^\top$ is the weight vector obtained from data. Two specific choices are $\hat{\mathbf{w}}_{\text{MMA}}$ and $\hat{\mathbf{w}}_{\text{JMA}}$ obtained from MMA and JMA, respectively.

5. Simulation Studies

In this section, we assess the numerical performance of our scalable model averaging using three simulation experiments. First, we consider a case when the full model is correct. Note that this case violates Condition C.2. We aim to investigate the performance of our scalable model averaging when some theoretical conditions do not hold. Second, we consider a case that all candidate models are misspecified. Third, we evaluate the performance of the proposed method with high-dimensional data. We compare five model averaging methods: (a) smooth

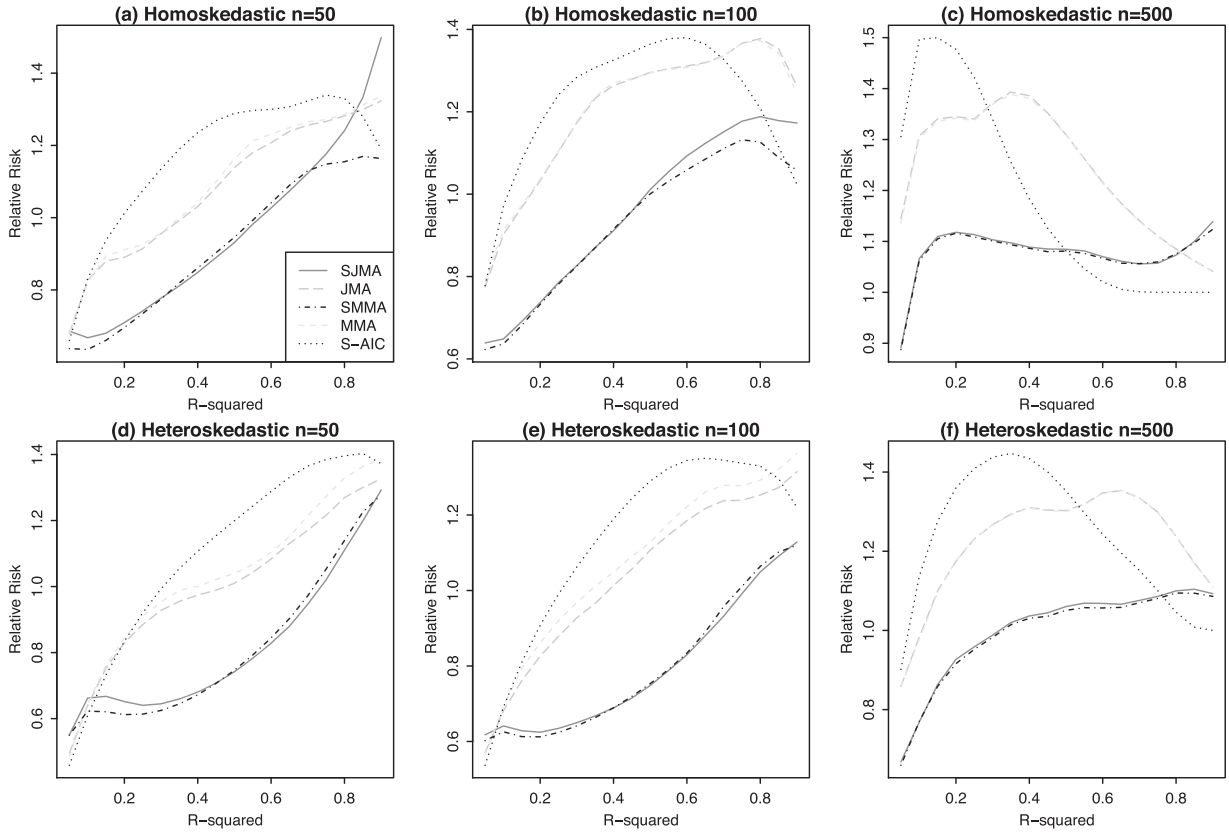

Figure 1. The risk performances in Example 1 with dimension $p = 5$. We normalize the risk by dividing it by the risk of the full model.

AIC (S-AIC) in Buckland et al. (1997), (b) Mallows' model averaging (MMA) in Hansen (2007), (c) Jackknife model averaging (JMA) in Hansen and Racine (2012), (d) our scalable Mallows' model averaging (SMMA), and (e) our scalable Jackknife model averaging (SJMA). In the high-dimensional setting, we compare our scalable model averaging estimators with the approach proposed by Ando and Li (2014) as well as the high-dimensional model selection method. Their performance is evaluated in terms of the risk under the squared loss function $L = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$.

Example 1 (Full model is correct). Data are generated from the model $y_i = \sum_j^p \theta_j x_{ji} + \epsilon_i$. Rows of the n by p predictor matrix $\mathbf{X}$ is generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ with the (i, j)th element $\sigma_{ij} = \rho^{|i-j|}$. The value of ρ is set to 0.6. The sample size n is set to 50 (small sample size), 100 (moderate sample size), and 500 (large sample size), and the dimension is set to $p = 5$. The true coefficients $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ are generated from a uniform distribution of $(0, 2)$ and then fixed. For the errors ϵ_i, two settings are considered: one is homoscedastic, with $\epsilon_i \sim N(0, \sigma^2)$; the other is heteroscedastic, with $\epsilon_i \sim 0.5N(0, \sigma^2) + 0.5N(0, 3\sigma^2)$, which means that for half data ϵ_i's are from $N(0, \sigma^2)$ and for the other half ϵ_i's are from $N(0, 3\sigma^2)$. The value of σ^2 is selected to control the R^2 to vary on a grind between 0.1 and 0.9, in which the R^2 is defined as $R^2 = \frac{n^{-1}\sum(\mu_i - \bar{\mu})^2}{\sigma^2 + n^{-1}\sum_{i=1}^n (\mu_i - \bar{\mu})^2}$ with $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$.

To evaluate each estimator, we compute the empirical risk (expected squared error) by calculating the average loss across $B = 1000$ simulations. Each risk is rescaled by the risk of the full model. We present the results in Figure 1. From these figures, it is seen that, for both homoscedastic and heteroscedastic settings, SMMA and SJMA are much better than MMA and JMA for most R^2 values. Meanwhile, the performances of SMMA and SJMA get closer as n increases. The difference between the scalable and original model averaging estimators gets smaller as R^2 increases, and the scalable estimators sometimes are slightly worse than the original ones when R^2 is very high, because when R^2 is very high (especially when n is large as well), the original JMA and MMA methods can reach the corresponding optimal risk easily. In this situation, our method can also easily reach its optimal risk, and both risks are close to each other, as shown in Theorem 1.

We also consider the case that covariates contain both discrete and continuous variables to check the impact of different types predictor distributions. We put the results in Figure A.1 of the supplementary materials for similarity and saving space.

Computational cost. The computational times (in seconds) form different model averaging estimators with various values of p are reported in Table 1. From the table, the computational costs of JMA and MMA are growing exponentially as the predictor dimension p increases, and when $p > 13$, the computer could not handle the optimization due to the size of the weighting vectors. Even for S-AIC, the computer cannot handle the calculation with dimension $p > 20$ due to estimating so many candidate models. However, for our scalable methods, SJMA and SMMA, the required computational times are very short (close to 0 sec). Even when $n = 500$ and $p = 50$, the required

Table 1. Computation costs (seconds) of different frequentist model averaging estimators with different feature dimension p.

p	5	7	9	10	11	12	13	15	20	25	50
						$n = 50$					
S-AIC	0.000	0.022	0.054	0.070	0.163	0.366	0.821	5.95	2345	–	–
JMA	0.000	0.040	1.72	17.44	114.5	856	6696	–	–	–	–
MMA	0.000	0.043	1.393	13.32	87.5	712	6030	–	–	–	–
SJMA	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
SMMA	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	–
						$n = 500$					
S-AIC	0.007	0.026	0.108	0.226	0.690	1.336	2.022	9.15	3078	–	–
JMA	0.174	0.734	5.088	24.94	145.8	1113	7880	–	–	–	–
MMA	0.108	0.551	4.059	17.92	107.0	802	6864	–	–	–	–
SJMA	0.044	0.046	0.057	0.062	0.074	0.085	0.100	0.133	0.140	0.176	0.346
SMMA	0.028	0.030	0.030	0.032	0.035	0.038	0.041	0.046	0.072	0.104	0.172

NOTE: The computation is not affordable for JMA and MMA when $p > 13$ and for S-AIC when $p > 20$. The optimization is performed by "solve.QP" function from "quadprog" package in R language on a Macbook Pro with 3 GHz intel i7 processor and 8 GB memory running OS X operation system. "–" means that the calculation is computationally infeasible.

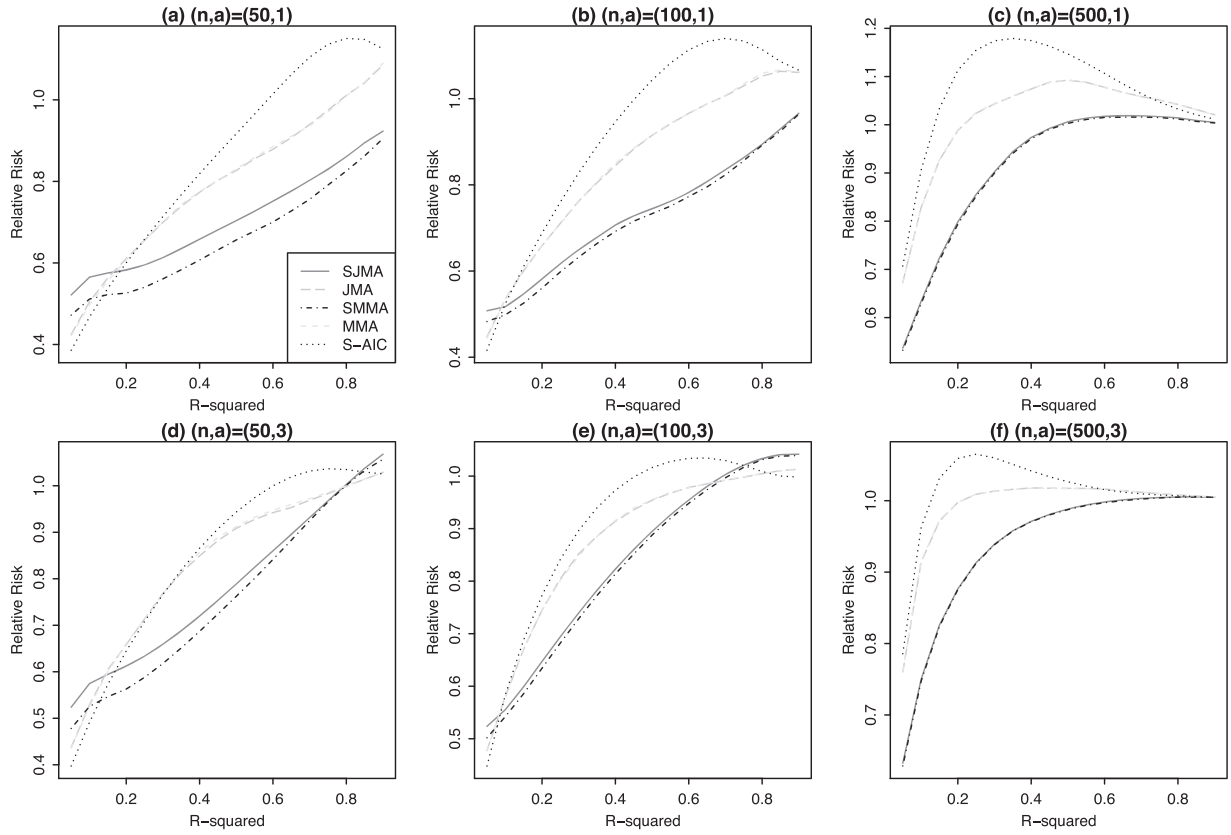

Figure 2. The risk performance of Example 2. The homoscedastic setting is considered here. We normalize the risk by dividing it by the risk of the maximal model. Note the R^2 here is also normalized because of the existing modeling bias.

times of SJMA and SMMA are less than 0.35 sec and 0.18 sec, respectively.

Example 2 (*Model misspecification*). In this example, we consider the case of model misspecification, that is, all candidate models are wrong models. The model setup in this example is similar to that in Example 1, except that the true model used to generate data contains an additional variable. To be specific, data are generated from a linear model with $p = 9$ using the same setup as in Example 1. However, only the first eight predictors are used in candidate model construction and the ninth variable is not used in any working model. The coefficient a of the ninth variable is set to be 1 and 3 to represent different degrees of misspecification. Results for the homoscedastic setting are presented in Figure 2. The results for the heteroscedastic setting are similar and thus are omitted to save space. We normalize the risk by dividing it by the risk of the maximal model. From the figure, SMMA and SJMA are much better than MMA and JMA. As a increases, the advantages of SMMA and SJMA become less significant, but they are still uniformly better than MMA and JMA, respectively. These observations show that our scalable methods have good performance for misspecified models.

Example 3 (*High-dimensional data*). In this example, we consider the case of high-dimensional data, and compare our scalable model averaging method with the approach proposed by

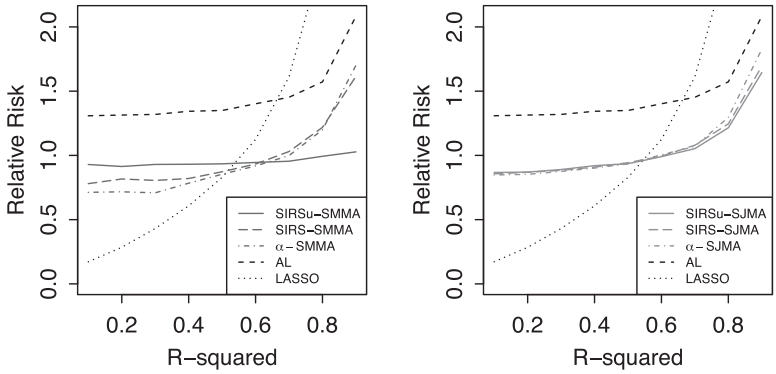

Figure 3. Example 3: The risk performances. We normalize the risk by dividing it by the risk of the full model. We split the results into two figures for better presentation.

Ando and Li (2014) (AL). We implement the three approaches discussed in Section 4. For the SJMA and SMMA after screening $\mathbf{X}$ by the SIRS, we refer them as SIRS-SJMA and SIRS-SMMA, respectively; for the SJMA and SMMA after screening $\mathbf{U}$ by the SIRS, we refer them as SIRSu-SJMA and SIRSu-SMMA, respectively; for the SJMA and SMMA by removing left singular vectors with small singular values, we refer them as α-SJMA and α-SMMA, respectively. Here α stands for the fact that we keep the singular vectors such that the summation of their singular values is $100\alpha\%$ of the summation of all singular values. We also consider the performance of the high-dimensional model selection methods, LASSO, MCP (Zhang 2010), and SCAD (Fan and Li 2001). The results for these three methods are similar so we report the results for the LASSO and omit results for the other two. We use R package program "ncvreg" (Breheny and Huang 2011) to implement the LASSO, and perform 5-fold cross-validation to select the penalty parameter.

We adopt the same linear model setting used in Ando and Li (2014). To be specific, $p = 2000$ predictors are generated from the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ with the (i,j)th element $\sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.6$. Among the $p = 2000$ predictors, 50 of them are active predictors (with nonzero regression coefficients), and they are spaced evenly. That is, the true predictors are X_j for $j = 40(s-1)+1$ with $s = 1, 2, \ldots, 50$. The coefficients θ_j's for the true predictors are generated from a normal distribution with mean 0 and standard deviation 0.5. The sample size n is set to 50. Ando and Li (2014) fixed the variance of the error term to be $\sigma^2 = 0.2^2$. Different from theirs, we choose multiple values of σ^2 to let R^2 vary on a grind between 0.1 and 0.9.

In this example, for α-SMMA and α-SJMA, $\alpha = 0.99$ is used so that the sum of singular values for singular vectors to use is about 99% of the sum of all singular values. The number of predictors by SIRS is set to $k = 48$. We show the performance of α-SMMA, α-SJMA, SIRS-SMMA, SIRS-SJMA, SIRSu-SMMA, SIRSu-SJMA, and AL in term of risk for different values of R^2 in Figure 3. It is seen that the risks of all scalable model averaging estimators are much smaller than those of AL for all cases. Meanwhile, our scalable model averaging also outperforms the LASSO as $R^2 > 0.5$. It indicates that our scalable model averaging outperforms the AL approach in the high-dimensional setting, and screening $\mathbf{X}$ or $\mathbf{U}$, and dropping left singular vectors with small singular values are effective strategies to apply scalable model averaging in high-dimensional data. In addition, from the perspective of the computational costs, the AL approach, on average, takes 0.5 sec in this high-dimensional setting, while our scalable approach, on average, just takes 0.024 sec. Therefore, we observe from this simulated high-dimensional experiment that our scalable model averaging methods has desirable performance for high-dimensional data. In addition, for JMA, the strategy by screening $\mathbf{U}$ performs very similar to that by screening $\mathbf{X}$. For MMA, this strategy outperforms that by screening $\mathbf{X}$ when R^2 is large, and vice versa.

6. Real Data Examples

In this section, we analyze two real datasets to further evaluate the performances of our scalable model averaging method for both the $p < n$ and $p > n$ settings. The two datasets considered here are the engineer wage dataset (ENGIN) and the firm-level data dataset (CEOSAL2) in Wooldridge (2003). The performances of the proposed methods are evaluated on the testing set in terms of the average prediction error under the squared loss function $L = n_t^{-1}\|\hat{\boldsymbol{\mu}} - \mathbf{y}\|^2$, where n_t is the size of the testing set, and $\hat{\boldsymbol{\mu}}$ and $\mathbf{y}$ are predicted values and observed values in the testing set, respectively.

Example 4 (ENGIN). The response variable is the log of monthly salary ("lwage"), and the features include dummy variables male, highgrad, college, grad, polytech, highdrop, and non-dummy variables educ, swage, exper, pexper, expersq, lswage, pexpersq, mleeduc, and mleeduc0. The intercept term is added in the model, so the dimension of the predictor vector is $p = 16$. Totally, there 403 observations in the dataset.

We consider five model averaging methods, S-AIC, MMA, JMA, SMMA, and SJMA. To evaluate the prediction performance of these estimators, we randomly select a training set of sample size $n = 350$ for model fitting and use the rest of $n_t = 403 - n$ observations as testing data. We repeat this process 100 times to obtain the average performances. Although the dimension of the predictor vector is only moderate ($p = 16$), it is still computationally infeasible to obtain the weights for MMA and JMA without the nested-model assumption. Thus, to implement the MMA and JMA methods, candidate models are constructed with the nested structure by correlation ranking suggested in Ando and Li (2014).

Table 2. Performances of model averaging estimators for the CEOSAL2 data.

n	α=95%		α=99%		SIRS		SIRSu		AL	LASSO
	α-SJMA	α-SMMA	α-SJMA	α-SMMA	SJMA	SMMA	SJMA	SMMA		
100	310.9	311.4	319.7	313.5	314.8	312.2	319.7	313.8	330.7	329.1
120	366.1	366.2	374.2	370.0	373.3	369.9	374.2	370.2	389.5	393.5
150	274.9	274.7	281.0	275.0	278.9	274.9	281.0	275.1	289.2	297.2

NOTE: We randomly split the dataset as a training set of size n and a testing set of $n_t = 177 - n$ observations. The prediction errors are the arithmetic means from 100 replications.

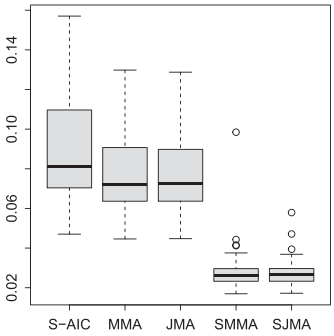

Figure 4. Performances of model averaging estimators for the ENGIN dataset. We randomly split the dataset as a training set of size $n = 350$ and a testing set of $n_t = 403 - n$ observations. The losses are summarized from 100 replications.

Figure 4 reports the results of these model averaging estimators by plotting the prediction errors under the squared loss in boxplots. It shows that our scalable model averaging estimators, SMMA and SJMA, are much better than the existing model averaging estimators. This example suggests that the scalable model averaging estimators are more efficient in prediction for moderate dimension.

Example 5 (CEOSAL2). We use the CEOSAL2 dataset to evaluate the performance of our scalable model averaging estimators in high-dimensional setting. The CEOSAL2 dataset is a dataset about firm's profits. The dependent variable is the profits as % of sales, and the predictors include dummy variables collage grad, and non-dummy variables salary, age, comten, centen, sales, mktval, lsalary, lsales, lmktval, comtensq, and cetensq. The original dataset contains 177 observations and 13 raw features. We add lsalary^2, lsales^2 and lmktval^2 to extend the number of predictors. We use the 16 raw features to create $p = 121$ features by adding interaction terms and an intercept and removing the almost dependent terms (the correlation more than 0.995).

We compare our scalable model averaging estimators with AL (Ando and Li 2014), and the high-dimensional model selection method, LASSO. Note we implement MCP, SCAD as well and omit the results here because of their similarity to that of the LASSO. For this dataset, we consider the training data of sample size $n = 100, 120, 150$ and use the rest $n_t = 177-n$ observations as the testing data. We repeat this process 100 times to obtain the average performance. We adopt the three strategies discussed in Section 4. For α-SJMA and α-SMMA, we set α to be 0.95 and 0.99. For the SIR screening, we set the number of variables to keep by the SIRS to be k that corresponds to $\alpha = 0.99$. The performances of the scalable model averaging estimators are reported in Table 2.

From the table, the scalable model averaging estimators produce smaller prediction errors than AL and LASSO in all cases. It suggests that the scalable model averaging estimators are more efficient in estimation for high-dimensional data. It seems sufficient to set $\alpha \geq 0.95$. We also observe that both strategies of screening $\mathbf{X}$ and screening $\mathbf{U}$ are efficient. Therefore, similar to the simulated experiment, this real data example also shows the applicability of the scalable model averaging method to high-dimensional data.

7. Conclusion and Discussion

In this article, we have proposed a scalable frequentist model averaging method. This method is computationally efficient, as it only needs to average p candidate models instead of 2^p candidate models. Moreover, the computational efficiency is achieved *without* sacrificing the prediction efficiency. Actually, the empirical prediction efficiency is even improved due to the decreased size of weights. We have rigorously proved that the minimum loss of the scalable model averaging estimator is asymptotically equal to that for the traditional model averaging estimator, which may involve averaging up to 2^p candidate models. We further have established the asymptotic optimality of the scalable Mallows/Jackknife model averaging estimators. It is worthy to mention that our scalable model averaging estimators can easily be applied into high-dimensional data. Compared with existing methods, the advantages of our method in terms of computation and estimation efficiency are also demonstrated via extensive simulations and real data examples. Specifically speaking, the empirical results indicate that our method not only saves much computational time, but also produces more accurate predictions. For the high-dimensional case, a comparison with Ando and Li's (2014) approach indicates that our method achieves a higher estimation efficiency with remarkably lower computational cost.

There are important questions about the scalable model averaging method that remain for future research. In high-dimensional setting, we propose three procedures to apply our scalable model averaging: screening the covariates, discarding singular vectors with small singular values, and screening the singular vectors. Empirically, these procedures have desirable performance. However, theoretically, whether the minimum loss from these approaches are equal to the minimum loss from the traditional model averaging with all candidate models is unclear. Ando and Li's investigation has this theoretical limitation as well, because the asymptotic optimality of their approach was established based on grouped variables rather than on considering all possible candidate models. This is a research topic for future investigation. Another interesting

topic is how to extend the idea to more general settings such as generalized linear models and other nonlinear models.

Supplementary Materials

The supplementary materials contains the proofs of Theorems 1 and 2 and additional simulation studies in Section 5.

Acknowledgments

The authors thank the Editor Professor Ivan Canay, the Associate Editor, and two reviewers for their insightful comments and suggestions that significantly improve the article.

Funding

Zhu's work was partially supported by NNSF of China grants 11301514 and 71532013, and by Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01) and 111 Project (No.B18015). Zhang's research was supported by NNSF grant (71925007, 72091212 and 12288201) and the CAS Project for Young Scientists in Basic Research (YSBR-008) Wang was partially supported by NSF grant 2105571.

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. (B. N. Petrov and F. Cźaki), pp. 267–281, Budapest: Akademiai Kaidó. [1]

Ando, T. and Li, K. (2014), "A Model-Averaging Approach for High-Dimensional Regression," *Journal of American Statistical Association*, 109, 254–265. [2,5,6,8,9]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [5]

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection," *Annals of Applied Statistics*, 5, 232–253. [8]

Buckland, S., Burnham, K., and Augustin, N. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [1,6]

Charkhi, A., Claeskens, G., and Hansen, B. E. (2016), "Minimum Mean Squared Error Model Averaging in Likelihood Models," *Statistica Sinica*, 26, 809–840. [2]

Claeskens, G. (2012), "Focused Estimation and Model Averaging with Penalization Methods, an Overview," *Statistica Neerlandica*, 66, 272–287. [2]

Clyde, M., Desimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Noconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [8]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," (with Discussion), *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [5]

Hansen, B. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [1,2,4,5,6]

Hansen, B., and Racine, J. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [1,2,4,5,6]

Hjort, N., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of American Statistical Association*, 98, 879–899. [1]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [1]

Jolliffe, I. T. (1982), "A Note on the Use of Principal Components in Regression," *Journal of the Royal Statistical Society*, Series C, 31, 300–303. [2]

Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [1]

Liang, H., Zou, G., Wan, A., and Zhang, X. (2011), "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of American Statistical Association*, 106, 1053–1066. [1]

Magnus, J., and Durbin, J. (1999), "Estimation of Regression Coefficients of Interest When other Regression Coefficients are of no Interest," *Econometrica*, 67, 639–643. [2]

Park, S. H. (1981), "Collinearity and Optimal Restrictions on Regression Parameters for Estimating Responses," *Technometrics*, 23, 289–295. [2]

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of American Statistical Association*, 92, 179–191. [1]

Wooldridge, D. (2003), *Introductory Econometrics*, Mason, OH: Thompson South-Western. [1,8]

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of American Statistical Association*, 96, 574–588. [1]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of American Statistical Association*, 100, 1202–1214. [1]

Zhang, C. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [8]

Zhu, L., Li, L., Li, R., and Zhu, L. (2011), "Model-Free Feature Screening for Ultrahigh-Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [5]

Zhu, L.-P. (2011), "Extending the Scope of Inverse Regression Methods in Sufficient Dimension Reduction," *Communications in Statistics. Theory and Methods*, 40, 84–95. [2]