Grouped Model Averaging for Finite Sample Size

Xinyu Zhang and Aman Ullah*
Chinese Academy of Sciences and University of California, Riverside

Abstract: This paper studies grouped model averaging methods for finite sample size situation. Sufficient conditions under which the grouped model averaging estimator dominates the ordinary least squares estimator are provided. A class of grouped model averaging estimators, g-class, is introduced, and its dominance condition over the ordinary least squares is established. All theoretical findings are verified by simulation examples. We also apply the methods to the analysis of the grain output data of China.

KEY WORDS: Finite Sample Size, Mean Squared Error, Model Averaging, Sufficient Condition.

JEL Classification codes: C13, C21.

1 Introduction

Over the past two decades, there has been a substantial amount of interest in model averaging. Within the Bayesian paradigm, model averaging has long been a popular approach; see, for example, Hoeting et al. (1999) for a comprehensive review. In recent years, within the frequentest paradigm, model averaging methods has been proposed, including weighting strategies using scores of information criteria (Buckland et al., 1997; Claeskens et al., 2006; Hjort and Claeskens, 2003; Zhang and Liang, 2011; Zhang et al., 2012), asymptotically optimal methods (Hansen, 2007; Hansen and Racine, 2012; Liang et al., 2011; Liu and Okui, 2013), plug-in model averaging (Liu, 2015), model averaging marginal regression (Li et al., 2015), among others. Frequentist model averaging technique has also been utilized in many contexts such as constructing optimal instruments (Kuersteiner and Okui, 2010), autoregressive models (Hansen, 2010), mixed-effects models (Zhang et al., 2014), factor augmented regression models (Cheng and Hansen, 2015), and quantile regression models (Lu and Su, 2015), see Ullah and Wang (2013) for a recent review.

It is well known that the estimation based on a "small" model can be more efficient than that based on a "large" model, but the former one can lead to substantial biases. Model averaging

*Corresponding Author: E-mail addresses: aman.ullah@ucr.edu.

aims to a trade-off between efficiency and biases. However, in most of the existing literature, model averaging methods were generally compared asymptotically, or by simulation performance, and there was no analytical finite sample study on the condition under which a model averaging estimator dominates the ordinary least squares (OLS) estimator with respect to mean squared error (MSE).

Recently, Hansen (2014) developed grouped model averaging methods, in which the regressors are firstly grouped in sets and then a model averaging method is implemented based on these sets. Assume there are M groups of regressors and let k_m denote the size of the m^{th} group. He proved that when the condition $k_m \geq 4$ for $m = 2, \ldots, M$ is satisfied, the asymptotic MSE (i.e., the MSE depending on an asymptotic distribution) of the grouped Mallows model averaging (GMMA) estimator is globally smaller than that of the OLS estimator. This is a very inspiring result because the condition is very simple and does not depend on any unknown parameter. However, his result is asymptotic and it is based on the assumption of local mis-specification in which some coefficients are of order $n^{-1/2}$ where n is the sample size. This is a useful procedure, although it also draws criticism because of its realism; see, for example, the discussions in Ishwaran and Rao (2003) and Raftery and Zheng (2003). In Hansen (2014), although the asymptotic theory was developed under the local mis-specification assumption, to make simulation experiments correspond to actual econometric practice, the author sets the coefficients to be fixed.

The main contribution of this paper is to develop new results on the exact dominance of the grouped model average estimators over the OLS estimator. In developing these results, local mis-specification assumption is not used. Also, the results are exact in the sense that they are valid for any sample size, especially when the sample size is small. For example, in China, most of annual data began in 1978 when the reform and opening-up policies were launched. When the sample size tends to infinity we show that Hansen's (2014) result reduces as a special case of our exact results. Also, our results show that for the finite sample situation, Hansen's (2014) asymptotic dominance condition $k_m \geq 4$ for $m = 2, \ldots, M$ is not sufficient. In view of this, based on a slight modification of Mallows' criterion, a class of grouped model averaging estimators, g-class, is then introduced , and it is shown that a member of this class has the same exact dominance condition

over the OLS as the Hansen's (2014) asymptotic dominance condition. Furthermore, we apply the group model averaging methods in analysis of the grain output data of China, which has only 35 observations.

The remainder of this paper is organized as follows. Section 2 introduces some estimators and basic theoretic results. Section 3 presents the MSE comparison between the GMMA estimator (and its modified versions) and the OLS estimator, and provides the sufficient conditions under which these grouped model averaging estimators dominate the OLS estimator. A g-class grouped model averaging estimator is also presented. Sections 4 and 5 provide simulation examples and a real data analysis, respectively. Section 6 concludes the paper. Technical proofs are contained in an Appendix.

2 Estimation

We are concerned with a linear regression model:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2), \qquad i = 1, \ldots, n \tag{1}$$

where y_i is a scalar dependent variable, $\mathbf{x}_i(p \times 1)$ are independent variables, $\boldsymbol{\beta}(p \times 1)$ is a coefficient vector, e_i is an error term, and $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$ are assumed to be independent. To simplify notation we treat the independent variables as fixed, but the theory applies also to random independent variables if proper conditions are imposed. In matrix notation, the model (1) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. We assume that $\mathbf{X}$ has full column rank $p < n$. The OLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \sim \text{Normal}\left\{\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}. \tag{3}$$

The variance σ^2 is estimated by $\widehat{\sigma}^2 = (n-p)^{-1}\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{y}\|^2$, where $\|\cdot\|^2$ stands for the Euclidean norm. It is well known that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$ are independent and

$$(n-p)\widehat{\sigma}^2\sigma^{-2} \sim \mathcal{X}^2(n-p), \qquad E(\widehat{\sigma}^2) = \sigma^2, \qquad var(\widehat{\sigma}^2) = 2(n-p)^{-1}\sigma^4. \tag{4}$$

Suppose that we have M groups of regressors. We combine M nested sub-models of (2) candidate models, and the m^{th} candidate model includes the first m groups of variables of $\mathbf{X}$, denoted by $\mathbf{X}_m$. Denote the group size of the m^{th} group by k_m. Let $v_m = \sum_{j=1}^{m} k_j$, and thus v_m is the number of variables used in the m^{th} candidate model and is also the number of columns of $\mathbf{X}_m$.

Let $\mathbf{\Pi}_m$ be a selection matrix so that $\mathbf{\Pi}_m = (\mathbf{I}_{v_m}, \mathbf{0}_{v_m \times (p-v_m)})$ and thus $\mathbf{X}_m = \mathbf{X}\mathbf{\Pi}_m^{\mathrm{T}}$. Define a $p \times p$ matrix $\mathbf{A}_m = \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m)^{-1}\mathbf{\Pi}_m(\mathbf{X}^{\mathrm{T}}\mathbf{X})$. Under the m^{th} candidate model, the restricted OLS estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_m &= \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m)^{-1}\mathbf{X}_m^{\mathrm{T}}\mathbf{y} \\
&= \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m)^{-1}\mathbf{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&= \mathbf{\Pi}_m^{\mathrm{T}}(\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m)^{-1}\mathbf{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} \\
&= \mathbf{A}_m\widehat{\boldsymbol{\beta}}_{\text{OLS}}. \qquad (5)
\end{aligned}$$

For the M^{th} candidate model, $\widehat{\boldsymbol{\beta}}_M = \widehat{\boldsymbol{\beta}}_{\text{OLS}}$. The grouped model averaging estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m\widehat{\boldsymbol{\beta}}_m,$$

where w_m is the weight corresponding to the m^{th} candidate model and $\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$, belonging to weight set $\mathcal{H} = \left\{\mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\right\}$.

Let $\mathbf{v} = (v_1, \ldots, v_M)^{\mathrm{T}}$. Hansen (2007) proposed choosing weights by minimizing Mallows' criterion

$$C(\mathbf{w}) = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \mathbf{y}\right\|^2 + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v}. \qquad (6)$$

Let

$$\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_M)^{\mathrm{T}} = \arg\min_{\mathbf{w}\in\mathcal{H}} C(\mathbf{w}),$$

so that the combined estimator $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ is the grouped Mallows model averaging (GMMA) estimator of $\boldsymbol{\beta}$.

Next, similar to Hansen (2014), we define cumulative weights

$$w_m^* = w_1 + \cdots + w_m$$

and $\mathbf{w}^* = (w_1^*, \ldots, w_M^*)^{\mathrm{T}}$. Then, $w_m = w_m^* - w_{m-1}^*$ for $m \geq 2$, $w_1 = w_1^*$, and $\mathbf{w} \in \mathcal{H}$ is equivalent to

$$\mathbf{w}^* \in \mathcal{H}^* \equiv \left\{ \mathbf{w}^* \in [0,1]^M : 0 \leq w_1^* \leq \cdots \leq w_M^* = 1 \right\} \tag{7}$$

and the grouped model averaging estimator $\widehat{\boldsymbol{\beta}}(\mathbf{w})$ can be rewritten as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\mathbf{w}) &= \sum_{m=2}^{M} (w_m^* - w_{m-1}^*) \widehat{\boldsymbol{\beta}}_m + w_1^* \widehat{\boldsymbol{\beta}}_1 \\ &= \sum_{m=1}^{M} w_m^* \widehat{\boldsymbol{\beta}}_m - \sum_{m=1}^{M-1} w_m^* \widehat{\boldsymbol{\beta}}_{m+1} \\ &= \widehat{\boldsymbol{\beta}}_{\text{OLS}} - \sum_{m=1}^{M-1} w_m^* (\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m). \end{aligned} \tag{8}$$

Let $\widehat{w}_m^* = \widehat{w}_1 + \cdots + \widehat{w}_m$, $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, \ldots, \widehat{w}_M^*)^{\mathrm{T}}$,

$$b_m = \left\| \mathbf{X} \widehat{\boldsymbol{\beta}}_m \right\|^2, \tag{9}$$

and

$$C^*(\mathbf{w}^*) = \sum_{m=1}^{M-1} \left\{ {w_m^*}^2 (b_m - b_{m+1}) - 2\widehat{\sigma}^2 w_m^* k_{m+1} \right\}. \tag{10}$$

From Lemma 1 of Hansen (2014), we have

$$C(\mathbf{w}) = C^*(\mathbf{w}^*) + b_M + 2\widehat{\sigma}^2 v_M \tag{11}$$

and

$$\widehat{\mathbf{w}}^* = \arg\min_{\mathbf{w}^* \in \mathcal{H}^*} C^*(\mathbf{w}^*). \tag{12}$$

Hence, from (5) and (9)-(12), we know that both $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{w}}^*$ depend on $\mathbf{y}$ through $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$.

Since weights $\widehat{w}_1, \ldots, \widehat{w}_M$ are determined by data, the indexes of candidate models with positive weights are random. We use $\left\{ m_1(\mathbf{y}), \ldots, m_{J(\mathbf{y})}(\mathbf{y}) \right\}$ to denote the indexes set, where $J(\mathbf{y})$ and $m_{j(\mathbf{y})}(\mathbf{y})$ depend on $\mathbf{y}$. By the analysis of the above paragraph, we know that $C(\mathbf{w})$ depends on $\mathbf{y}$ through $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$, so we can write $J(\mathbf{y})$ and $m_{j(\mathbf{y})}(\mathbf{y})$ as $J(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)$ and $m_{j(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)$.

Instead, for simplicity, we write them as J and m_j, although they are random. Thus, $w^*_{m_j} = \cdots = w^*_{m_{j+1}-1}$ and $w^*_{m_J} = 1$. From the proof of Theorem 1 of Hansen (2014), we have

$$\begin{aligned} C^*(\mathbf{w}^*) &= \sum_{m=1}^{M-1} \left\{ {w^*_m}^2 (b_m - b_{m+1}) - 2\widehat{\sigma}^2 w^*_m k_{m+1} \right\} \\ &= \sum_{j=1}^{J-1} \sum_{\ell=m_j}^{m_{j+1}-1} \left\{ w^{*2}_\ell (b_{\ell+1} - b_\ell) - 2\widehat{\sigma}^2 w^*_\ell k_{\ell+1} \right\} + \sum_{\ell=m_J}^{M-1} \left\{ w^{*2}_\ell (b_{\ell+1} - b_\ell) - 2\widehat{\sigma}^2 w^*_\ell k_{\ell+1} \right\} \\ &= \sum_{j=1}^{J-1} \left\{ {w^*_{m_j}}^2 (b_{m_{j+1}} - b_{m_j}) - 2\widehat{\sigma}^2 w^*_{m_j} (v_{m_{j+1}} - v_{m_j}) \right\} + (b_M - b_{m_J}) - 2\widehat{\sigma}^2 (v_M - v_{m_J}), \end{aligned}$$

which is minimized by

$$\widehat{w}^*_{m_j} = \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{b_{m_{j+1}} - b_{m_j}}, \qquad j = 1, \ldots, J-1, \tag{13}$$

when $w^* \in \mathcal{H}^*$ (see (7) for the definition of $\mathcal{H}^*$).

3 MSE Comparison

3.1 MSE of the GMMA estimator

Let

$$\begin{aligned} q(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2) &= I(m_J < M) \left\{ 2\sigma^2 (v_M - v_{m_J}) - \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J} \right\|^2 \right\} \\ &\quad + \sigma^4 \sum_{j=1}^{J-1} \frac{\left\{ (n-p-2)(n-p)^{-1} (v_{m_{j+1}} - v_{m_j}) - 4 \right\} (v_{m_{j+1}} - v_{m_j})}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2}, \end{aligned}$$

where $I(\cdot)$ denotes the indicator function as usual. For any estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, its MSE is defined by $E \left\| \mathbf{X}\widetilde{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta} \right\|^2$.

Theorem 1. $E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta} \right\|^2 = \sigma^2 p - E \left\{ q(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2) \right\}$.

See Appendix A.1 for the proof of Theorem 1. From Theorem 1, we have the following result.

Corollary 1. *If* $(n-p-2)(n-p)^{-1} k_m \geq 4$ *for all* $m \geq 2$, *then*

$$E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta} \right\|^2 < E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} \right\|^2,$$

i.e., $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ *dominates* $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.

See Appendix A.2 for the proof of Corollary 1. We note that the result in Corollary 1 provides the exact dominance condition for the GMMA estimator over the OLS estimator in the MSE sense. Corollary 3 of Hansen (2014) is a special case of Corollary 1 above when n tends to infinity, i.e., a sufficient condition for the GMMA estimator dominating (asymptotically) the OLS estimator is $k_m \geq 4$ for all $m \geq 2$. Irrespective of sample size, our Corollary 1 indicates that there is a scale $(n-p-2)(n-p)^{-1} < 1$ associated with k_m and when $(n-p-2)(n-p)^{-1}k_m \geq 4$ for all $m \geq 2$, the GMMA estimator dominates the OLS estimator.

3.2 MSE of g-Class Grouped Model Averaging Estimators

In the Mallows' criterion (6), the first term measures the model fit, while the second term measures the model complexity and serves as a penalty, where the constant 2 can be viewed as a tuning parameter. To be more general, we consider weight choice criterion as follows:

$$\widetilde{C}(\mathbf{w}, g) = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \mathbf{y}\right\|^2 + 2g\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v},$$

where the tuning parameter 2 is multiplied by a positive constant g. Obviously,

$$\widetilde{C}(\mathbf{w}, 1) = C(\mathbf{w}).$$

Let $\widetilde{\mathbf{w}}_g = \arg\min_{\mathbf{w}\in\mathcal{H}} \widetilde{C}(\mathbf{w}, g)$, $\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_g)$ be the g-class grouped model averaging estimator which is equal to $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ (GMMA estimator) for $g = 1$, i.e., $\widetilde{\mathbf{w}}_1 = \widehat{\mathbf{w}}$. Define

$$\begin{aligned}\widetilde{q}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2) &= I(m_J < M)\left\{2\sigma^2(v_M - v_{m_J}) - \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\} \\ &\quad + \sigma^4\sum_{j=1}^{J-1}\frac{g\left[\left\{2 - g(n-p+2)(n-p)^{-1}\right\}(v_{m_{j+1}} - v_{m_j}) - 4\right](v_{m_{j+1}} - v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}.\end{aligned}$$

Theorem 2. $E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_g) - \mathbf{X}\boldsymbol{\beta}\right\|^2 = \sigma^2 p - E\left\{\widetilde{q}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)\right\}$.

See Appendix A.3 for the proof of Theorem 2. From Theorem 2 and the proof of Corollary 1, it is straightforward to obtain the following results.

Corollary 2. *If $0 < g \leq 2(n-p)(n-p+2)^{-1}(k_m - 2)k_m^{-1}$ for all $m \geq 2$, then $\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_g)$ dominates $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.*

Some special cases of g are described below.

Corollary 3. *When* $g = (n-p+2)^{-1}(n-p)$, $\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_g)$ *dominates* $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ *given that* $k_m \geq 4$ *for all* $m \geq 2$.

Motivated by Corollary 3, we define a new model averaging method with weight vector

$$\widetilde{\mathbf{w}}_{g=(n-p+2)^{-1}(n-p)}.$$

This method dominates the OLS under the condition that $k_m \geq 4$ for all $m \geq 2$ is satisfied, which is free from the sample size and number of regressors. Since it is a modified version of GMMA, we term it mGMMA.

Recently, Zhang et al. (2015) proposed choosing weights by minimizing the following Kullback-Leibler criterion

$$\text{KL}(\mathbf{w}) = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \mathbf{y}\right\|^2 + 2(n-p)(n-p-2)^{-1}\widehat{\sigma}^2\mathbf{w}^{\text{T}}\mathbf{v},$$

so $\widetilde{C}\left\{\mathbf{w}, (n-p)(n-p-2)^{-1}\right\} = \text{KL}(\mathbf{w})$. Define $\widehat{\mathbf{w}}_{\text{KL}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \text{KL}(\mathbf{w})$. This method is called grouped Kullback-Leibler model averaging (GKLMA), and it is a member of g-class grouped model averaging estimators with $g = (n-p)(n-p-2)^{-1}$. From Corollary 2, it is straightforward to obtain the following result.

Corollary 4. *When* $(n-p-6)(n-p-2)^{-1}k_m \geq 4$ *for all* $m \geq 2$, $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{KL}})$ *dominates* $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.

We have observed above that for special cases of $g = 1$, $g = (n-p)(n-p+2)^{-1}$, and $g = (n-p)(n-p-2)^{-1}$ we get GMMA, mGMMA, and GKLMA estimators, respectively. It will be an interesting topic to find the optimum value of g, in the g-class grouped model averaging estimator, for which MSE is minimum. However, this is extremely challenging and out of scope of this paper

4 Simulation Examples

In this section, we use simulation examples to verify the theoretical results of the previous section. Specifically, we should have the following findings:

Finding I. When $(n-p-2)(n-p)^{-1}k_m \geq 4$ and $(n-p-6)(n-p-2)^{-1}k_m \geq 4$ for all $m \geq 2$, the GMMA and GKLMA will always yield smaller MSEs than the OLS.

Finding II. When $(n-p-2)(n-p)^{-1}k_m < 4$ for any $m \geq 2$, the GMMA can perform worse than the OLS; when $(n-p-6)(n-p-2)^{-1}k_m < 4$ for any $m \geq 2$, the GKLMA can perform worse than the OLS.

Finding III. When $k_m \geq 4$ for $m \geq 2$, the mGMMA will always yield smaller MSEs than the OLS.

Findings I-II will verify Corollaries 1 and 4. Finding III will verify Corollary 3.

The simulation setting is from Hansen (2014); that is

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ji} + e_i, \qquad i = 1, \ldots, n$$

with $e_i \sim \text{Normal}(0,1)$ $(i = 1, \ldots, n)$, $x_{ji} \sim \text{Normal}(0,1)$, $\beta_0 = 0$, $\beta_j = cj^{-\alpha}$ $(j = 1, \ldots, p-1)$, and $\alpha \in \{0, 1, 2, 3\}$. The coefficient c is selected to vary the population R^2 in $\{0.1, 0.2, \ldots, 0.9, 0.98\}$. We use the following configurations of n, p and k_m:

I. $n = 12$, $p = 5$, $k_1 = 1$, $k_2 = 4$;
II. $n = 16$, $p = 9$, $k_1 = 1$, $k_2 = 4$, $k_3 = 4$;
III. $n = 30$, $p = 6$, $k_1 = 1$, $k_2 = 5$;
IV. $n = 35$, $p = 11$, $k_1 = 1$, $k_2 = 5$, $k_3 = 5$.

All MSEs in estimating $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^{\mathrm{T}}$ are calculated by using 10,000 replications. The MSEs of model averaging methods are normalized by that of the OLS estimator, so a MSE below one indicates that the estimator has smaller MSE than the OLS. Figures 1-4 show the MSEs for $\alpha = 0, 1, 2, 3$, respectively.

In Configurations III-IV, $(n-p-2)(n-p)^{-1}k_2 = 4.583 \geq 4$ and $(n-p-6)(n-p-2)^{-1}k_2 = 4.091 \geq 4$. It is seen from bottom two panels of Figures 1-4 that the GMMA and GKLMA always lead to smaller MSE than the OLS. This is Finding I.

In Configurations I-II, $(n-p-2)(n-p)^{-1}k_2 = 3.667 < 4$ and $(n-p-6)(n-p-2)^{-1}k_2 = 3.273 < 4$. It is seen from top two panels of Figures 1-4 that the GMMA and GKLMA sometimes lead to larger MSE than the OLS. This is Finding II.

Figures 1-4 show that the mGMMA always lead to smaller MSE than the OLS. This is Finding III.

In addition, we find that for all grouped model averaging methods, their MSEs can be much lower than that of the OLS, especially when R^2 is small, i.e., residual variance is high. This finding is encouraging in view of the fact that R^2 is often small in many cross sectional models.

5 Analysis of Real Data

In China, annul data are often very short, most of which begin in 1978, or even later. Hence in this section, we applied the grouped model averaging methods that have good statistical properties under finite sample size case to the analysis of the grain output (tons) data of China.

The data consists of the annul observations from 1978 to 2012 from National Bureau of Statistics of China at http://www.stats.gov.cn. The grain includes rice, wheat, corn bean, and tubers. The logarithm grain output (GO) is shown in Figure 5. Four independent variables collected are sown area of grain crops (SAGC) (hectares), employed persons (EP) (persons), total agricultural machinery power (TAMP) (kw), and consumption of chemical fertilizer (CCF) (tons). Figure 6 illustrates logarithm of these variables. We used a linear regression model

$$\Delta\log(GO_i) = \beta_1 + \beta_2\Delta\log(SAGC_i) + \beta_3\Delta\log(TAMP_i) + \beta_4\Delta\log(CCF_i) + \beta_5\Delta\log(EP_i) + e_i,$$

for $i = 1979, \ldots, 2012$, where $\Delta A_i = A_i - A_{i-1}$. When implementing the grouped model averaging methods, we let the intercept as a group and the remaining variables as a group, so $k_1 = 1$ and $k_2 = 4$ in this case.

Table 1 shows the estimates by the OLS and the three grouped model averaging methods, where the standard errors of the grouped model averaging estimates are obtained by bootstrap. It is seen that all methods indicate that the sown area of grain crops (SAGC) has positive impact on the grain output. We further evaluated these methods by an out-sample prediction. We used observations before 2001 to estimate parameters (so the sample size is 23) and use the remaining observations to calculate mean squared prediction errors (MSPE). Table 2 presents the results. It is seen that all model averaging methods perform similarly and better than the OLS. This performance is reasonable in view of the facts that the adjusted-Rsquare in the estimation is 0.453 and our simulation

results show that the grouped model averaging methods perform much better than the OLS when R^2 is small.

6 Concluding Remarks

Firstly we have developed new results on deriving the condition under which the GMMA estimator dominates the OLS estimator in the exact MSE sense. This exact condition depends on the sample size and the number of regressors. In a special case, when n tends to infinity, we have shown that the exact dominance condition reduces to the condition derived by Hansen (2014) based on an asymptotic MSE. This condition is free from sample size and the number of regressors. The g-class grouped model averaging estimator is also introduced and its exact dominance condition is obtained, which depends on the sample size and number of regressors. It is shown that a member of this class has an exact dominance condition free from the sample size and number of regressors, and it is the same as the asymptotic dominance condition of Hansen (2014). Secondly we remark that as Hansen (2014), our theory is also confined to the context of nested models. Extension of the current analysis to non-nested models will be very challenging. Thirdly, the MSE comparison of the current paper is built under the normally distributed and homoscedastic error. Developing MSE comparison under other error cases is also an interesting topic for future research. Lastly , the grouped model averaging is based on a pre-supposed grouping structure. The existing grouping procedures such as octagonal shrinkage and clustering algorithm for regression (Bondell and Reich, 2008) may be utilized in applications.

Acknowledgment

Zhang's research was supported by National Natural Science Foundation of China (Grant no. 71101141 and 11271355). Also, Ullah's research was supported by the Academic Senate Grant at UCR.

Appendices

A.1 Proof of Theorem 1

Note that the set $\{m_1, \ldots, m_J\}$ which contains indexes of candidate models with positive weights is random and depends on $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$. When $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$ vary, the set $\{m_1, \ldots, m_J\}$ can also vary, but it is a piecewise constant function of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$ and is almost differentiable in the sense of Stein (1981) except for a finite number of points. Hence, in the following proof, when taking derivatives with respect to $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$, we take $\{m_1, \ldots, m_J\}$ be a constant set.

Since model m_j is nested within model m_{j+1}, it is easily to obtain the following results

$$(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})^{\mathrm{T}}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{M} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) = 0 \tag{A.1}$$

and

$$b_{m_{j+1}} - b_{m_j} = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}\right\|^2 - \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2 = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2. \tag{A.2}$$

It follows from (3), (5)-(13), (A.1)-(A.2), Stein Lemma (Stein, 1981), and the independence between $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\widehat{\sigma}^2$ that

$$\begin{aligned}
&E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta}\right\|^2 \\
&= E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{m=1}^{M-1}\widehat{w}^*_m(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m})\right\|^2 \\
&= E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1}\sum_{\ell=m_j}^{m_{j+1}-1}\widehat{w}^*_\ell(\mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell}) - I(m_J < M)\sum_{\ell=m_J}^{M-1}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell})\right\|^2 \\
&= E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1}\widehat{w}^*_{m_j}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_{M} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J})\right\|^2 \\
&= E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1}\frac{\widehat{\sigma}^2(v_{m_{j+1}} - v_{m_j})}{b_{m_{j+1}} - b_{m_j}}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})\right. \\
&\quad \left. - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_{M} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J})\right\|^2 \\
&= E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1}\frac{\widehat{\sigma}^2(v_{m_{j+1}} - v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})\right. \\
&\quad \left. - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_{M} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J})\right\|^2 \\
&= \sigma^2 p + E\sum_{j=1}^{J-1}\frac{\widehat{\sigma}^4(v_{m_{j+1}} - v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2} + E\left\{I(m_J < M)\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{M} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&-2E\left[\left\{I(m_J<M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_M-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J})\right\}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta})\right]\\
&-2E\left[\left\{\sum_{j=1}^{J-1}\frac{\widehat{\sigma}^2(v_{m_{j+1}}-v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})\right\}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta})\right]\\
&=\sigma^2p+E\left\{\widehat{\sigma}^4\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\}+E\left\{I(m_J<M)\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\}\\
&-2\sigma^2E\left[I(m_J<M)\mathrm{trace}\left\{(\mathbf{A}_M-\mathbf{A}_{m_J})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}\right]\\
&-2E\left(\widehat{\sigma}^2E\left[\left\{\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})\right\}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta})|\widehat{\sigma}^2\right]\right)\\
&=\sigma^2p+E\left\{\widehat{\sigma}^4\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\}\\
&+E\left[I(m_J<M)\left\{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2-2\sigma^2(v_M-v_{m_J})\right\}\right]\\
&-2\sigma^2E\left[\widehat{\sigma}^2\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\mathrm{trace}\left\{(\mathbf{A}_{m_{j+1}}-\mathbf{A}_{m_j})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}\right]\\
&+4\sigma^2E\sum_{j=1}^{J-1}\left\{\widehat{\sigma}^2\frac{(v_{m_{j+1}}-v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^4}\right\}\\
&\quad\times\mathrm{trace}\left\{(\mathbf{A}_{m_{j+1}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}_{m_{j+1}}-\mathbf{A}_{m_j}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}_{m_j})\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}\\
&=\sigma^2p+E\left\{\widehat{\sigma}^4\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\}\\
&+E\left[I(m_J<M)\left\{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2-2\sigma^2(v_M-v_{m_J})\right\}\right]\\
&-2\sigma^2E\left\{\widehat{\sigma}^2\sum_{j=1}^{J-1}\frac{(v_{m_{j+1}}-v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\}+4\sigma^2E\sum_{j=1}^{J-1}\widehat{\sigma}^2\frac{v_{m_{j+1}}-v_{m_j}}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}. \qquad \text{(A.3)}
\end{aligned}
$$

Let $\widehat{a}=2^{-1}(n-p)\widehat{\sigma}^2\sigma^{-2}$ and $a=2^{-1}(n-p)$. From (4), we have

$$\widehat{a}\sim\text{Gamma}\left\{2^{-1}(n-p),1\right\} \qquad \text{(A.4)}$$

with mean a. So, by the independence between $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ and Lemma 2 of Shen and Huang

(2006), we obtain that for any constant c and any functions of $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$, $f_1(\widehat{\boldsymbol{\beta}}_{\text{OLS}}), \ldots, f_{J-1}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})$,

$$
\begin{aligned}
& E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&= E\left[E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^2 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&\quad + E\left[E\left\{(\widehat{\sigma}^2 - \sigma^2)\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^2 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \qquad \text{(A.5)}
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&= E\left[E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&\quad + 4\sigma^4(n-p)^{-2} E\left[E\left\{(\widehat{a} - a)(\widehat{a} + a)\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&\quad + 4\sigma^4(n-p)^{-2} E\left[E\left\{\widehat{a}\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} \\
&\quad + 2\sigma^2(n-p)^{-1} E\left[E\left\{\widehat{\sigma}^2\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})|\widehat{\boldsymbol{\beta}}_{\text{OLS}}\right\}\right] \\
&= \sigma^4\left\{1 + 2(n-p)^{-1}\right\} E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}. \qquad \text{(A.6)}
\end{aligned}
$$

In addition,

$$
2(v_{m_{j+1}} - v_{m_j})^2 - 4(v_{m_{j+1}} - v_{m_j}) - \left\{1 + 2(n-p)^{-1}\right\}(v_{m_{j+1}} - v_{m_j})^2
$$

$$= (v_{m_{j+1}} - v_{m_j}) \left\{ 2(v_{m_{j+1}} - v_{m_j}) - 4 - (n-p+2)(n-p)^{-1}(v_{m_{j+1}} - v_{m_j}) \right\}$$
$$= (v_{m_{j+1}} - v_{m_j}) \left\{ (n-p-2)(n-p)^{-1}(v_{m_{j+1}} - v_{m_j}) - 4 \right\}. \quad \text{(A.7)}$$

The result of Theorem 1 is implied by the above (A.3)-(A.7) and the definition of $q(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)$.

A.2 Proof of Corollary 1

From (A.4) and Lemma 2 of Shen and Huang (2006), we have

$$E\left\{(\widehat{\sigma}^2 - \sigma^2) I(m_J < M)(v_M - v_{m_J})\right\} = 0. \quad \text{(A.8)}$$

Similar to the derivation of (28) in Hansen (2014), we can obtain that if $m_J < M$, then

$$b_M - b_{m_J} \leq \widehat{\sigma}^2 (v_M - v_{m_J}). \quad \text{(A.9)}$$

It is seen from (A.2), (A.8)-(A.9) and the definition of $q(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)$ that when $(n-p-2)(n-p)^{-1}k_m \geq 4$ for all $m \geq 2$, we have

$$\begin{aligned} E\left\{q(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)\right\} &\geq E\left[I(m_J < M)\left\{2\sigma^2(v_M - v_{m_J}) - \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\}\right] \\ &= E\left[I(m_J < M)\left\{\widehat{\sigma}^2(v_M - v_{m_J}) - b_M - b_{m_J}\right\}\right] \\ &\quad - E\left\{I(m_J < M)(\widehat{\sigma}^2 - \sigma^2)(v_M - v_{m_J})\right\} \\ &\quad + \sigma^2 E\left\{I(m_J < M)(v_M - v_{m_J})\right\} \\ &\geq \sigma^2 E\left\{I(m_J < M)(v_M - v_{m_J})\right\}. \end{aligned} \quad \text{(A.10)}$$

When $m_J < M$, $I(m_J < M)(v_M - v_{m_J})$ is larger than zero, which, along with (A.10), implies the result of Corollary 1.

A.3 Proof of Theorem 2

By using the same steps of (A.5) and (A.6), we have

$$E\left\{g\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} = g\sigma^2 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}$$

and

$$E\left\{g^2\widehat{\sigma}^4 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}$$
$$= g^2\sigma^4\left\{1 + 2(n-p)^{-1}\right\} E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}.$$

In addition,

$$
\begin{aligned}
&2g(v_{m_{j+1}} - v_{m_j})^2 - 4g(v_{m_{j+1}} - v_{m_j}) - g^2\left\{1 + 2(n-p)^{-1}\right\}(v_{m_{j+1}} - v_{m_j})^2 \\
&= g(v_{m_{j+1}} - v_{m_j})\left[\left\{2 - g(n-p+2)(n-p)^{-1}\right\}(v_{m_{j+1}} - v_{m_j}) - 4\right].
\end{aligned}
$$

From above formulas, the definition of $\widetilde{q}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}, \widehat{\sigma}^2)$, and the proof of Theorem 1, we can obtain the result of Theorem 2.

References

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model selection: An integral part of inference," *Biometrics*, 53, 603–618.

Cheng, X. and Hansen, B. E. (2015), "Forecasting with factor-augmented regression: A frequentist model averaging approach," *Journal of Econometrics, forthcoming*.

Claeskens, G., Croux, C., and van Kerckhoven, J. (2006), "Variable selection for logistic regression using a prediction-focused information criterion," *Biometrics*, 62, 972–979.

Hansen, B. E. (2007), "Least squares model averaging," *Econometrica*, 75, 1175–1189.

— (2010), "Averaging estimators for autoregressions with a near unit root," *Journal of Econometrics*, 158, 142–155.

— (2014), "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, 5, 495–530.

Hansen, B. E. and Racine, J. (2012), "Jacknife model averaging," *Journal of Econometrics*, 167, 38–46.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: A tutorial," *Statistical Science*, 14, 382–417.

Ishwaran, R. H. and Rao, J. S. (2003), "Discussion," *Journal of the American Statistical Association*, 98, 922–925.

Kuersteiner, G. and Okui, R. (2010), "Constructing optimal instruments by first-stage prediction averaging," *Econometrica*, 78, 697–718.

Li, D., Linton, O., and Lu, Z. (2015), "A flexible semiparametric forecasting model for time series," *Universiyt of York, working paper*.

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011), "Optimal weight choice for frequentist model average estimators," *Journal of the American Statistical Association*, 106, 1053–1066.

Liu, C.-A. (2015), "Distribution theory of the least squares averaging estimator," *Journal of Econometrics, forthcoming*.

Liu, Q. and Okui, R. (2013), "Heteroskedasticity-robust Cp model averaging," *Econometrics Journal*, 16, 462–473.

Lu, X. and Su, L. (2015), "Jackknife model averaging for quantile regressions," *Journal of Econometrics, forthcoming*.

Raftery, A. E. and Zheng, Y. (2003), "Discussion: Performance of Bayesian model averaging," *Journal of the American Statistical Association*, 98, 931–938.

Shen, X. and Huang, H.-C. (2006), "Optimal model assessment, selection, and combination," *Journal of the American Statistical Association*, 101, 554–568.

Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 153, 1135–1151.

Ullah, A. and Wang, H. (2013), "Parametric and nonparametric frequentist model selection and model averaging," *Econometrics*, 1, 157–179.

Zhang, X. and Liang, H. (2011), "Focused information criterion and model averaging for generalized additive partial linear models," *Annals of Statistics*, 39, 174–200.

Zhang, X., Wan, A., and Zhou, S. 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.

Zhang, X., Zou, G., and Carroll, R. (2015), "Model averaging based on Kullback-Leibler distance," *Statistica Sinica, forthcoming*.

Zhang, X., Zou, G., and Liang, H. (2014), "Model averaging and weight choice in linear mixed-effects models," *Biometrika*, 101, 205–218.

Table 1: Estimates (Est.) and standard errors (s.e.) of coefficients in the real data model (14).

	OLS		GMMA		GKLMA		mGMMA	
Variables	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.e.
Intercept	0.015	0.018	0.016	0.023	-0.025	0.023	0.023	0.023
$\Delta\log(SAGC)$	1.682	0.328	1.503	0.341	0.782	0.343	1.430	0.340
$\Delta\log(TAMP)$	0.103	0.176	0.092	0.193	-0.344	0.192	0.185	0.194
$\Delta\log(CCF)$	-0.177	0.314	-0.158	0.358	0.507	0.356	-0.243	0.360
$\Delta\log(EP)$	0.357	0.157	0.319	0.187	0.073	0.186	0.169	0.188

Table 2: Mean squared prediction errors (MSPE) in prediction of the grain outputs from 2001 to 2012.

	OLS	GMMA	GKLMA	mGMMA
MSPE	6.659	5.969	5.927	6.010
s.e.	1.712	1.690	1.701	1.684

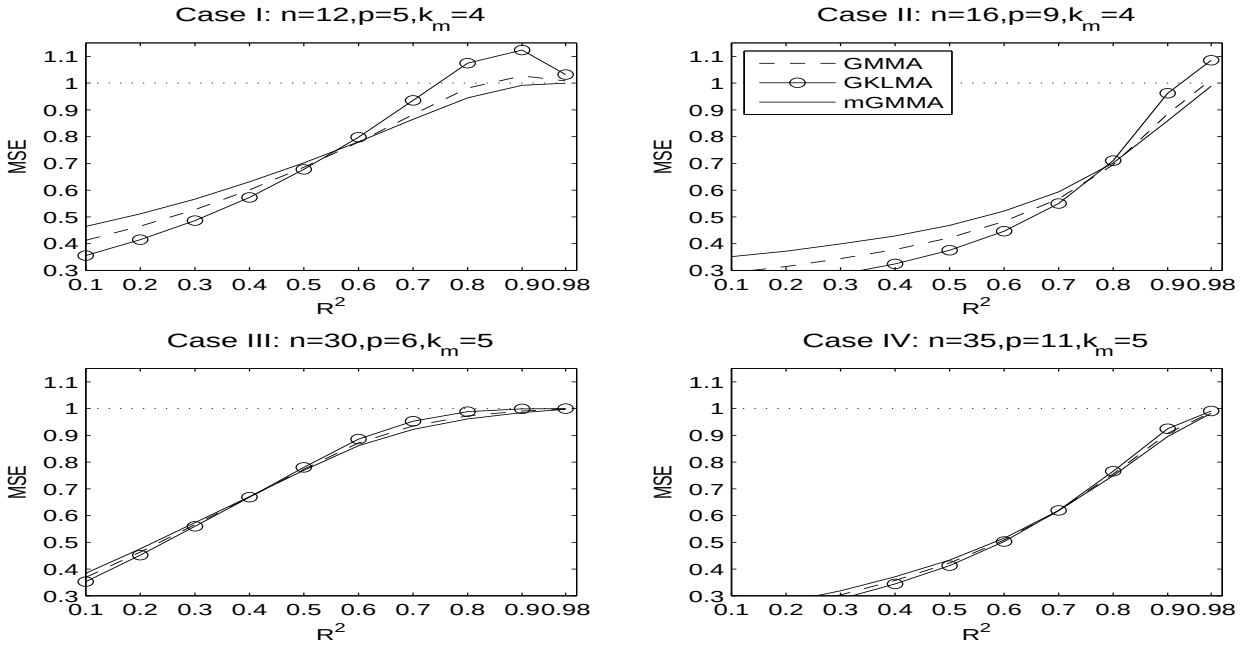

Figure 1: Simulation result: $\alpha = 0$

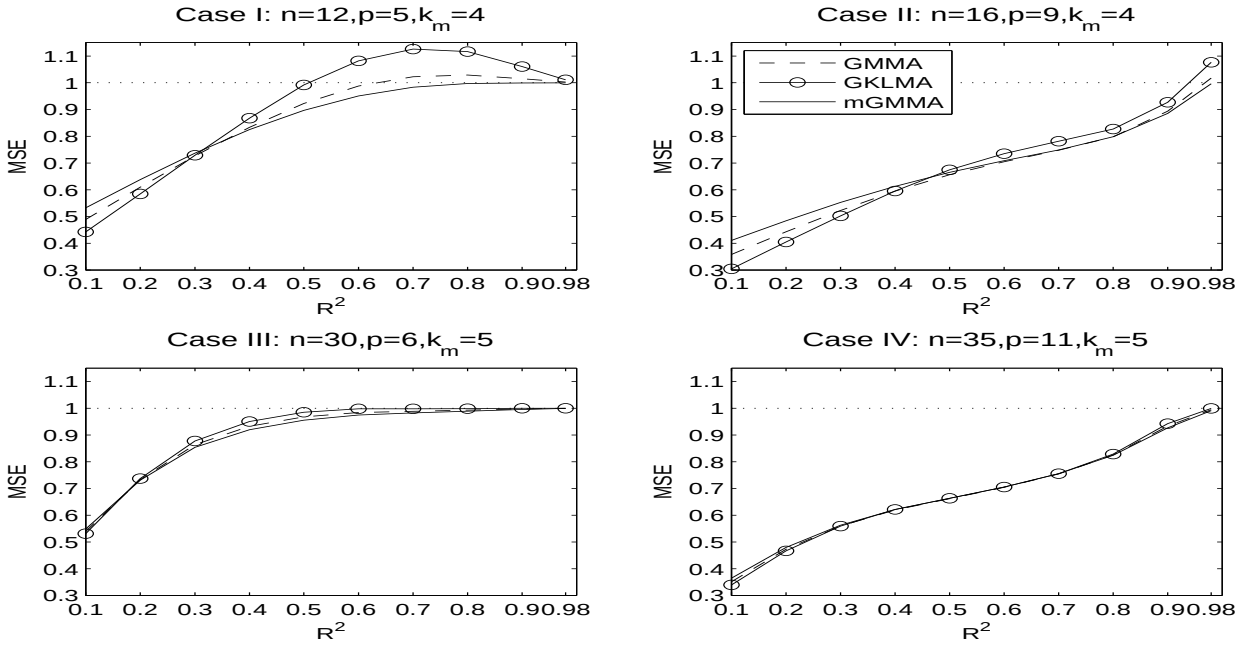

Figure 2: Simulation result: $\alpha = 1$

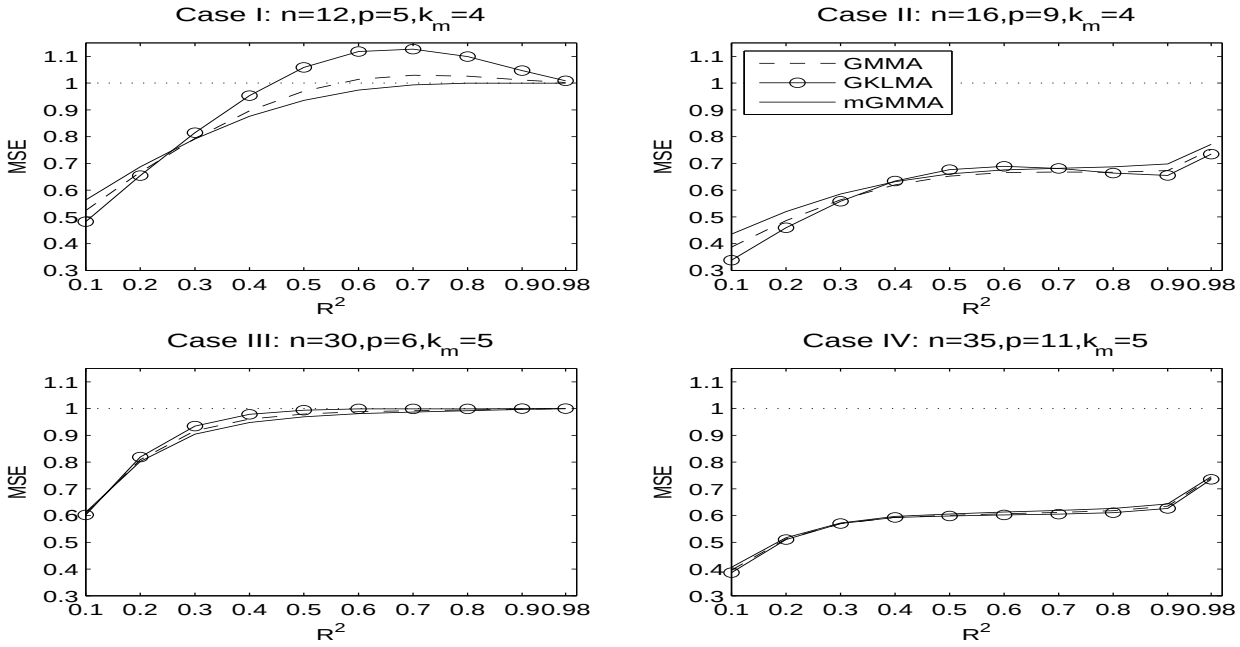

Figure 3: Simulation result: $\alpha = 2$

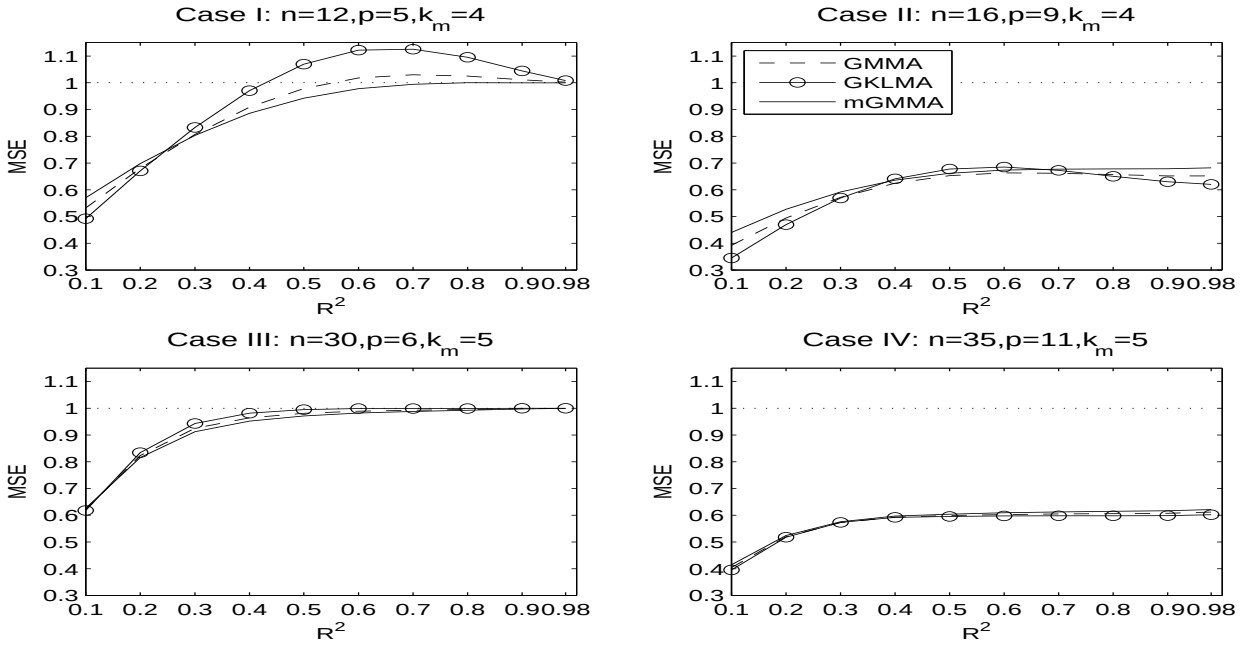

Figure 4: Simulation result: $\alpha = 3$

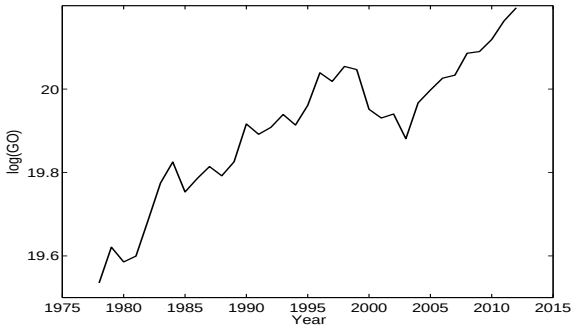

Figure 5: Dependent variable in application: logarithm of grain output (tons).

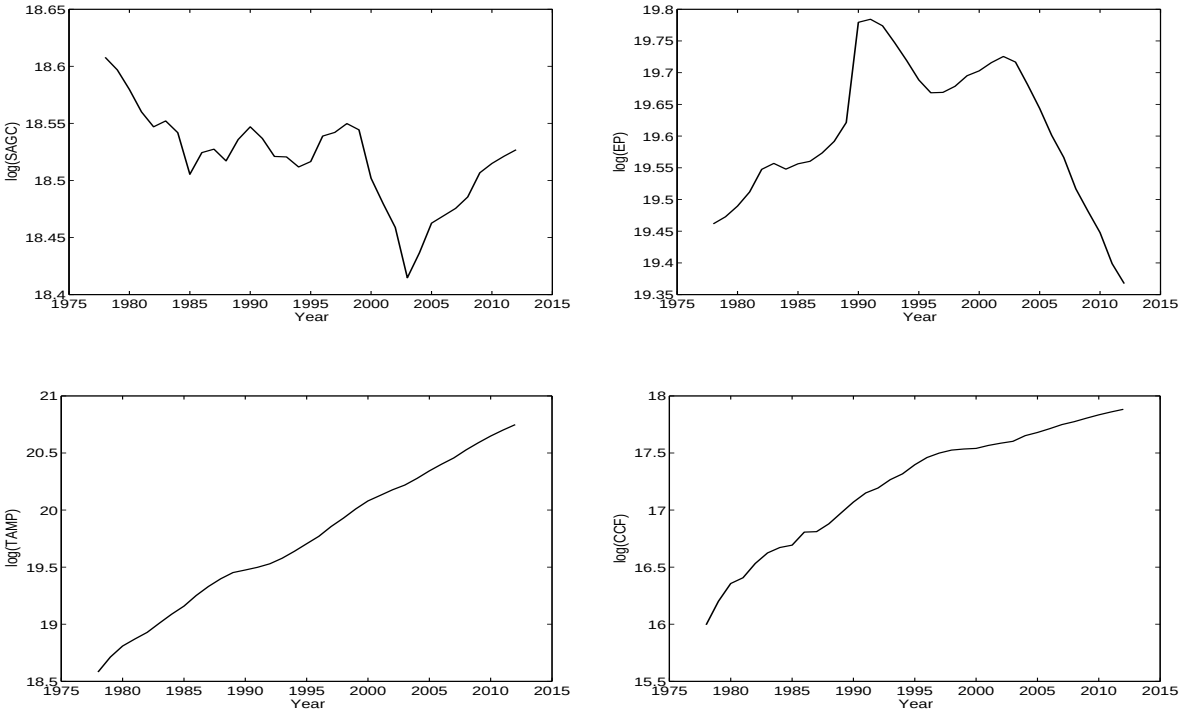

Figure 6: Independent variables in application. Top-left penal is logarithm of sown area of grain crops (SAGC) (hectares), top-right penal is logarithm of consumption of chemical fertilizer (CCF) (tons), bottom-left penal is logarithm of total agricultural machinery power (TAMP) (kw), and bottom-right penal is logarithm of employed persons (EP) (persons).