Statistica Sinica Preprint No: SS-2019-0230	
Title	Optimal Model Averaging Based on Generalized Method of Moments
Manuscript ID	SS-2019-0230
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0230
Complete List of Authors	Xinyu Zhang
Corresponding Author	Xinyu Zhang
E-mail	xinyu@amss.ac.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

OPTIMAL MODEL AVERAGING BASED ON GENERALIZED METHOD OF MOMENTS

Xinyu Zhang[1,2]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

[2] *Center for Forecasting Science, Chinese Academy of Sciences*

Abstract: **We propose a model averaging method that combines estimators from the generalized method of moments (GMM). Unlike other GMM-based model averaging procedures, this method allows all candidate models to be misspecified (not locally misspecified). We prove that when all candidate models are misspecified, the proposed method is optimal in the sense of minimizing the estimation loss; when there exists at least one correctly specified model, the method can achieve the common root-n convergence rate. Simulation experiments and an application to a housing market show the superiority of our method over other methods.**

Key words and phrases: Asymptotic optimality, Consistency, Generalized method of moments, Model averaging.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

1. Introduction

Model averaging and selection are the two main approaches used to deal with having many candidate models. Using model selection, we figuratively put all of our inferential eggs in one unevenly woven basket (Longford, 2005). Model averaging is a smoothed extension of model selection that substantially reduce the risk relative to that of selection (Hansen, 2014). Moreover, model averaging procedures can be more stable than those of model selection, for which a small change in the data can have a significant effect on the choice of the choice of model (Breiman, 1996; Yuan & Yang, 2005).

There are two types of model averaging: Bayesian model averaging (BMA) and frequentist model averaging (FMA). BMA has long been a popular statistical technique. Its main advantage is that inferences based on BMA are straightforward; see Hoeting et al. (1999) for a comprehensive review of this literature. FMA is commonly used to improve prediction or estimation precision. As discussed in Bates & Granger (1969) and Leung & Barron (2006), an average estimator often reduces the mean squared error (MSE) in an estimation. This is because it incorporates useful information from the relationship between the response and the covariates, providing a kind of insurance against selecting a very poor candidate model. Many

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

FMA methods have been proposed, including averaging weights based on the scores of information criteria (Buckland et al., 1997; Hjort & Claeskens, 2003, 2006; Zhang & Liang, 2011), optimal weighting (Hansen, 2007; Wan et al., 2010; Liang et al., 2011; Zhang et al., 2014; Zhang & Wang, 2019), adaptive weighting (Yang, 2001; Yuan & Yang, 2005; Zhang et al., 2013), plug-in methods (Liu, 2015; Yin et al., 2019), and model averaging marginal regression (Li et al., 2015; Chen et al., 2018). The optimal weighting method minimizes a weight choice criterion, and has been shown to provide the minimal prediction loss in a large sample sense. In the seminal work on optimal model averaging, Hansen (2007) combined the least squares estimators. Since then, a large body of literature has been formed on optimally combining least squares estimators or generalized least squares estimators, such as Hansen & Racine (2012), Liu & Okui (2013), Ando & Li (2014), Cheng & Hansen (2015), Liu et al. (2016), and Fang et al. (2019). Recently, optimal model averaging methods were extended to combine maximum likelihood estimators; see, for example, Zhang et al. (2016) and Ando & Li (2017). The weighted average least squares estimation is a method between BMA and FMA, using prior distributions and an analysis of the estimation risk from a frequentist perspective; see Magnus et al. (2010), Magnus et al. (2011), and De Luca et al. (2018).

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

In this study, we develop optimal model averaging based on the generalized method of moments (GMM). In general, the GMM is more applicable than the maximum likelihood method because the former only requires the moment functions, and does not require knowledge of the likelihood function. Despite the extensive literature on model averaging, few studies have explicitly examined GMM-based model averaging. Those that have include the works of DiTraglia (2016) and Cheng et al. (2019). DiTraglia (2016) combines GMM estimators from candidate models with different moment condition sets, and takes into account locally misspecified moment conditions. We describe the local misspecification in (2.4) of Section 2. Cheng et al. (2019) combines two GMM estimators, one of which is from a correctly specified candidate model. In contrast, we allow all candidate models to be misspecified (not locally misspecified).

To develop an optimal model averaging method for the GMM, following the classic model averaging literature, we propose a weight choice criterion by estimating the risk under the GMM framework. We prove that when all candidate models are misspecified, the corresponding model average estimator is optimal in the sense that it minimize the estimation loss. To provide more comprehensive support for using our method, we prove that it has root-n consistency when there are correctly specified candidate mod-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

els. Therefore, for a large sample sense, our method performs no worse than the commonly used methods that also achieve root-n consistency. In addition to providing theoretical justifications for the proposed method, we use a Monte Carlo study to demonstrate that the proposed averaging method outperforms the GMM and a selection method based on the GMM in a variety of settings, especially when the sample size is small.

The remainder of this paper is structured as follows. In Section 2, we introduce the candidate models and the GMM estimation. In Section 3, we introduce the proposed model average estimator based on the GMM. In Section 4, we show the asymptotic optimality and root-n consistency of the proposed method. In Sections 5 and 6, we report the results of a Monte Carlo study and a real-data application, respectively. Section 7 concludes the paper. The proofs of the theoretical results are given in the online Supplementary Material.

2. Candidate models and GMM estimation

Let $\boldsymbol{\theta}_{d\times 1}$ be an unknown vector, $\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta})_{p\times 1}$ be moments, and $\widehat{\boldsymbol{\mu}}_{p\times 1}$ be the sample moments. Thus, the moment conditions are

$$E\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta})\right\} = \mathbf{0}_{p\times 1}. \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Let $\boldsymbol{\mu}(\cdot)$ be the working moment function, which can be different from $\boldsymbol{\mu}_{\text{true}}(\cdot)$. As a result, the working moment conditions can be misspecified; that is,

$$E\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})\right\} \neq \mathbf{0}_{p\times 1}. \tag{2.2}$$

For example, when

$$y_i = X_{i1}\theta_1 + \cdots + X_{i(d-1)}\theta_{d-1} + \exp(X_{id}\theta_d) + \epsilon_i,$$

with $E(\epsilon_i|X_{i1},\ldots,X_{id}) = 0$, we have $\widehat{\boldsymbol{\mu}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$ and

$$\begin{aligned}\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}) \;=\; & \mathbf{X}^{\mathrm{T}}\Big[\left\{X_{11}\theta_1,\ldots,X_{1(d-1)}\theta_{d-1},\exp(X_{1d}\theta_d)\right\}^{\mathrm{T}}, \\ & \cdots,\left\{X_{n1}\theta_1,\ldots,X_{n(d-1)}\theta_{d-1},\exp(X_{nd}\theta_d)\right\}^{\mathrm{T}}\Big]^{\mathrm{T}},\end{aligned} \tag{2.3}$$

where $\mathbf{y} = (y_1,\ldots,y_n)^{\mathrm{T}}$ and $\mathbf{X} = \left\{(X_{11},\ldots,X_{1d})^{\mathrm{T}},\cdots,(X_{n1},\ldots,X_{nd})^{\mathrm{T}}\right\}^{\mathrm{T}}$. However, the working moment function may be incorrectly set as $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}$; that is, the function of the last variable X_{id} is misspecified. The local misspecification considered in DiTraglia (2016) is

$$E\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})\right\} = (\mathbf{0}^{\mathrm{T}}_{p_1\times 1}, \zeta^{\mathrm{T}}_{p_2\times 1}/\sqrt{n})^{\mathrm{T}}, \tag{2.4}$$

where ζ is an unknown vector and $p_1 + p_2 = p$. Notably, the setting in (2.2) is more general than that in (2.4).

Because we are uncertain whether some components of $\boldsymbol{\theta}$ should be set to zero, which determines whether certain variables should be used, we

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

consider M candidate models. For the mth candidate model, the unknown parameter vector is $\boldsymbol{\theta}_m$, which is a d_m-dimensional sub-vector of $\boldsymbol{\theta}$, such that $\boldsymbol{\theta}_m = \boldsymbol{\Pi}_m \boldsymbol{\theta}$, where $\boldsymbol{\Pi}_m$ is a projection matrix equal to $(\mathbf{I}_{d_m \times d_m}, \mathbf{0}_{d_m \times (d-d_m)})$, or a column permutation thereof. In the example following (2.2), when θ_d is very small, using $\boldsymbol{\mu}\{(\theta_1, \ldots, \theta_{d-1}, 0)^{\mathrm{T}}\}$ as the working moment function can be better than using $\boldsymbol{\mu}\{(\theta_1, \ldots, \theta_{d-1}, \theta_d)^{\mathrm{T}})\}$ in (2.5).

Under the mth candidate model, the GMM estimator of $\boldsymbol{\theta}_m$ is

$$\widehat{\boldsymbol{\theta}}_m = \mathrm{argmin}_{\boldsymbol{\theta}_m}[\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}} \boldsymbol{\theta}_m))^{\mathrm{T}} \boldsymbol{\Omega} (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}} \boldsymbol{\theta}_m)\}], \tag{2.5}$$

where $\boldsymbol{\Omega}$ is a positive-definite weighting matrix. Note that this is a special case of the classic minimum distance estimator and of the general estimator (Newey & McFadden, 1994), but not of a general GMM estimator in which the moment conditions are $E\left\{\boldsymbol{g}(\boldsymbol{\Pi}_m^{\mathrm{T}} \boldsymbol{\theta}_m)\right\} = \mathbf{0}_{p \times 1}$. Developing a model averaging method that combines the general GMM estimators is left to future research.

Note that the matrix $\boldsymbol{\Omega}$ and sample moments $\widehat{\boldsymbol{\mu}}$ do not vary with the model index m in (2.5), which implies that the candidate models use the same moment conditions. Hence, we combine models with different specifications in $\boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}} \boldsymbol{\theta}_m)$, rather than models with different moment conditions, as in DiTraglia (2016) and Cheng et al. (2019). We allow M and d_m to increase with the sample size n, but we need p to be unrelated to n. Note

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

that if d is large and all 2^d possible models are considered, then the computation burden will be very heavy. In this case, the model-screening methods developed in Ando & Li (2014) and Zhang et al. (2016) can be applied.

3. Model average estimator based on the GMM

Let $\mathbf{w} = (w_1, w_2, ..., w_M)^{\mathrm{T}}$ be a weight vector in the following set:

$$\mathcal{W} = \{\mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}. \tag{3.1}$$

We define the model average estimator of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}}(\mathbf{w}) \equiv \sum_{m=1}^{M} w_m \boldsymbol{\Pi}_m \widehat{\boldsymbol{\theta}}_m. \tag{3.2}$$

Because some components of the vectors $\boldsymbol{\Pi}_m \widehat{\boldsymbol{\theta}}_m$ are zeros, the model average estimator $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ is a type of shrinkage estimator, as pointed out by Liang et al. (2011) and Hansen (2014).

Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. A reasonable loss function to evaluate the model average estimator $\boldsymbol{\theta}(\mathbf{w})$ is

$$L(\mathbf{w}) \equiv [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)]^{\mathrm{T}} \boldsymbol{\Omega} [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)], \tag{3.3}$$

and the corresponding risk function is

$$R(\mathbf{w}) \equiv E\{L(\mathbf{w})\}. \tag{3.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Next, we propose a weight choice criterion by estimating the risk function $R(\mathbf{w})$. First, we list two conditions.

Condition (C.1) $\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)$ *satisfies the following central limit theorem:*

$$\sqrt{n}\left\{\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right\} \xrightarrow{d} \boldsymbol{\pi} \sim \mathit{Normal}(\mathbf{0}, \boldsymbol{V}),$$

where $\xrightarrow{d}$ *denotes convergence in distribution,* $\boldsymbol{\pi}$ *is a random vector, and* $\boldsymbol{V}$ *is a nonrandom positive-definite matrix.*

Condition (C.2) *For* $m \in \{1, \ldots, M\}$*, the derivatives* $\partial\boldsymbol{\mu}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ *and* $\partial\widehat{\boldsymbol{\theta}}_m/\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}$ *exist ar are continuous with respect to* $\boldsymbol{\theta}$ *and* $\widehat{\boldsymbol{\mu}}$*, respectively, and trace* $\left(\partial(\sqrt{n}\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\sqrt{n}\widehat{\boldsymbol{\mu}})/[\partial\sqrt{n}\left\{\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right\}^{\mathrm{T}}]\boldsymbol{\Omega}\boldsymbol{V}\right)$ *and* $\sqrt{n}\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})-\widehat{\boldsymbol{\mu}}\}\boldsymbol{\Omega}\sqrt{n}\{\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}$ *are uniformly integrable for* $\mathbf{w} \in \mathcal{W}$*.*

Condition (C.1) is the same as Assumption 1.9 of Harris & Mátyás (1999), where its rationality is discussed in detail. Condition (C.2) relates to the existence, continuity, and integrability. We propose the following weight choice criterion:

$$\begin{aligned}\widetilde{C}(\mathbf{w}) \;\equiv\; & [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\widehat{\boldsymbol{\mu}}]-2n^{-1}\mathrm{trace}\left(\boldsymbol{\Omega}\boldsymbol{V}\right)\\ & +2n^{-1}\mathrm{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}}\boldsymbol{\Pi}_m^{\mathrm{T}}\frac{\partial\widehat{\boldsymbol{\theta}}_m}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\boldsymbol{\Omega}\boldsymbol{V}\right]\\ & +\{\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}^{\mathrm{T}}\boldsymbol{\Omega}\{\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}. \qquad (3.5)\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Proposition 1 *Under Conditions (C.1)–(C.2), we have*

$$E\left\{\widetilde{C}(\mathbf{w})\right\} = R(\mathbf{w}) + o(n^{-1}). \tag{3.6}$$

The proof of Proposition 1 is given in Section S.1 of the Supplementary Material. The normal approximation is widely used in developing model selection criteria; see, for example, Hurvich & Tsai (1989). From (3.6), $\widetilde{C}(\mathbf{w})$ is an approximately unbiased estimator of the risk $R(\mathbf{w})$. By minimizing $\widetilde{C}(\mathbf{w})$ with respect to $\mathbf{w}$, the risk should also be minimized, but there are unknown parameters in $\widetilde{C}(\mathbf{w})$. Hence, the minimization is not feasible.

Let $\widehat{\boldsymbol{V}}$ be the preliminary estimator of $\boldsymbol{V}$. Andrews (1991) and Den Haan & Levin (1997) provide methods for estimating $\widehat{\boldsymbol{V}}$. Removing the terms unrelated to $\mathbf{w}$ and replacing $\boldsymbol{V}$ with its estimator, the criterion $\widetilde{C}(\mathbf{w})$ defined in (3.5) becomes

$$\begin{aligned} C(\mathbf{w}) \;\equiv\; & [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}] \\ & +2n^{-1}\mathrm{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \widehat{\boldsymbol{V}}\right], \end{aligned} \tag{3.7}$$

which can function as a weight choice criterion. By minimizing $C(\mathbf{w})$, we obtain the following weights:

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} C(\mathbf{w}). \tag{3.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

The first term of $C(\mathbf{w})$ measures the model fitness. To interpret the second term of $C(\mathbf{w})$, following Efron (2004), we define the degrees of freedom of the model average estimator $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ as

$$df(\mathbf{w}) = \operatorname{cov}\left\{\boldsymbol{\mu}^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}(\mathbf{w}))\boldsymbol{\Omega}^{1/2}, \widehat{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Omega}^{1/2}\right\}. \tag{3.9}$$

From the proof of Proposition 1, we know that the second term of $C(\mathbf{w})$ is an approximately unbiased estimator of the degrees of freedom $df(\mathbf{w})$. We refer to the resulting estimator $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}})$ the model average estimator based on the GMM ($\mathrm{MA}_{\mathrm{GMM}}$). When the weight components are restricted to one or zero, our method simplifies to a model selection method based on the GMM, called $\mathrm{MS}_{\mathrm{GMM}}$.

In general, the moment $\boldsymbol{\mu}(\boldsymbol{\theta})$ is an explicit function of $\boldsymbol{\theta}$; hence, the calculation of $\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}/\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}$ is straightforward. Next, we present a closed form for $\partial\widehat{\boldsymbol{\theta}}_m/\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}$. Write $\widehat{\boldsymbol{\theta}}_m = (\widehat{\theta}_{m,1}, \ldots, \widehat{\theta}_{m,d_m})^{\mathrm{T}}$. Let

$$\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m) = \frac{\partial\boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)^{\mathrm{T}}}{\partial\widehat{\boldsymbol{\theta}}_m}, \qquad \boldsymbol{A}_\tau(\widehat{\boldsymbol{\theta}}_m) = \frac{\partial\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)}{\partial\widehat{\theta}_{m,\tau}}, \tag{3.10}$$

$$\boldsymbol{D}_m = \left[\boldsymbol{A}_1(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\right\}, \ldots, \boldsymbol{A}_{d_m}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\right\}\right]_{d_m\times d_m}, \tag{3.11}$$

and

$$\boldsymbol{B}_m = \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\boldsymbol{A}^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_m), \tag{3.12}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

for $m = 1, \ldots, M$ and $\tau = 1, \ldots, d_m$, where d_m is the number of components in $\widehat{\boldsymbol{\theta}}_m$.

Proposition 2 *If Condition (C.2) holds, the derivatives $\partial \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)/\partial \widehat{\theta}_{m,\tau}$ for $m = 1, \ldots, M$ and $\tau = 1, \ldots, d_m$ exist, and the minimum singular value of the matrix $(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)$ is bounded away from a positive constant, for $m = 1, \ldots, M$, then*

$$\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} = -\left\{(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\right\}^{-1}(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}. \quad (3.13)$$

The proof of Proposition 2 is given in S.2 of the Supplementary Material. This proposition provides a closed form for the derivative $\partial \widehat{\boldsymbol{\theta}}_m/\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}$.

Remark 1 *When focusing on linear regression candidate models that have different regressor matrices, our criterion $C(\mathbf{w})$ simplifies to the Mallows' criterion introduced by Hansen (2007). Specifically, consider a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}|\mathbf{X} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{X}$ has a fixed full-column rank, and the regressor matrix for the mth candidate model is $\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}}$. Then, we have*

$$\widehat{\boldsymbol{\mu}} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{y}}{n},\ \boldsymbol{\mu}(\boldsymbol{\theta}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}}{n},\ \boldsymbol{\Omega} = \left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}\right)^{-1},\ \mathbf{V} = \sigma^2 E(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}), \quad (3.14)$$

where X_i^{T} is the ith row of $\mathbf{X}$. Let $\widehat{\sigma}^2$ be an estimator of σ^2. Then, $\widehat{\mathbf{V}} = \widehat{\sigma}^2\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$. From (3.10) and (3.11), we have

$$\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m) = \boldsymbol{\Pi}_m\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n},\ \boldsymbol{A}_\tau(\widehat{\boldsymbol{\theta}}_m) = \mathbf{0}_{d_m \times p},\ \boldsymbol{D}_m = \mathbf{0}_{d_m \times d_m}. \quad (3.15)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Hence, we can show that

$$trace\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \boldsymbol{\Omega} \widehat{\boldsymbol{V}}\right] = \widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m, \tag{3.16}$$

and thus

$$C(\mathbf{w}) = n^{-1}\|\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{y}\|^2 + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m - \mathbf{y}^{\mathrm{T}}\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right\}\mathbf{y}, \tag{3.17}$$

which is the Mallows' criterion in Hansen (2007) up to the term $\mathbf{y}^{\mathrm{T}}\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right\}\mathbf{y}$ *unrelated to* $\mathbf{w}$*. The proofs of (3.16) and (3.17) are provided in Section S.3 of the Supplementary Material.*

<u>Remark</u> 2 *In this remark, we consider linear regression models with instrumental variables. The linear regression model is still* $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$*, and there is an instrumental variable matrix* $\mathbf{Z}$ *that has a fixed full-column rank not smaller than that of* $\mathbf{X}$*, which also has a fixed full-column rank, and* $\boldsymbol{\epsilon}|\mathbf{Z} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n)$*. We fix* $\boldsymbol{\Omega} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n)^{-1}$*. For the* mth *candidate model, the regressor matrix is* $\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}}$*. Let* $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$ *and* $\widehat{\sigma}^2$ *be an estimator of* σ^2*. Then, we have*

$$\widehat{\boldsymbol{\mu}} = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{y}}{n},\ \boldsymbol{\mu}(\boldsymbol{\theta}) = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}}{n},\ \widehat{\mathbf{V}} = \widehat{\sigma}^2\frac{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}{n}, \tag{3.18}$$

$$\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m) = \boldsymbol{\Pi}_m \frac{\mathbf{X}^{\mathrm{T}}\mathbf{Z}}{n},\ \boldsymbol{A}_\tau(\widehat{\boldsymbol{\theta}}_m) = \mathbf{0}_{d_m \times p},\ \boldsymbol{D}_m = \mathbf{0}_{d_m \times d_m}. \tag{3.19}$$

Hence, similarly to (3.16) and (3.17), we can show that

$$trace\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \boldsymbol{\Omega} \widehat{\boldsymbol{V}}\right] = \widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m, \tag{3.20}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

and thus

$$C(\mathbf{w}) = n^{-1}\|\mathbf{P_Z}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{y}\|^2 + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m - \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P_Z})\mathbf{y}. \quad (3.21)$$

The proofs of (3.20) and (3.21) are given in Section S.3 of the Supplementary Material.

Lastly, note that if $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a linear function of $\boldsymbol{\theta}$ (i.e., there exists a matrix $\mathbf{Q}$ such that $\mu(\boldsymbol{\theta}) = \mathbf{Q}\boldsymbol{\theta}$), which is the case in the above remarks, then calculating $\widehat{\mathbf{w}}$ is extremely simple. Let $\widehat{\boldsymbol{g}}_m = \mathbf{Q}\widehat{\boldsymbol{\theta}}_m - \widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{G}} = (\widehat{\boldsymbol{g}}_1, \ldots, \widehat{\boldsymbol{g}}_M)$, and

$$\widetilde{\boldsymbol{g}} = \{\mathrm{trace}(\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Pi}_1^{\mathrm{T}}\partial\widehat{\boldsymbol{\theta}}_1/\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Omega}\widehat{\boldsymbol{V}}), \ldots, \mathrm{trace}(\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Pi}_M^{\mathrm{T}}\partial\widehat{\boldsymbol{\theta}}_M/\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Omega}\widehat{\boldsymbol{V}})\}^{\mathrm{T}}.$$

Then

$$C(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\widehat{\boldsymbol{G}}\mathbf{w} + 2n^{-1}\mathbf{w}^{\mathrm{T}}\widetilde{\boldsymbol{g}}. \quad (3.22)$$

Thus, the minimization of $C(\mathbf{w})$ with respect to $\mathbf{w}$ is simply a quadratic programming problem. Numerous software packages (e.g., quadprog of MATLAB) are available to solve this problem very efficiently even when M is very large.

4. Large-sample properties

In this section, we study the large-sample properties of the proposed $\mathrm{MA}_{\mathrm{GMM}}$ estimator $\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}})$. We first consider a common situation in which all candi-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

date models are misspecified (see Section 4.1 a the detailed description of the model misspecification). In that situation, we show that the estimator offers asymptotic optimality. Then, we consider an ideal situation in which at least one of the candidate models is correctly specified. In this case, the estimator is shown to have root-n consistency. All limiting processes discussed in this paper are as $n \to \infty$. The number of candidate models M can increase to infinity with n.

4.1 Asymptotic optimality under misspecified candidate models

When no value of $\boldsymbol{\theta}_m$ exists such that $\boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\boldsymbol{\theta}_m) = \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)$, we say that the mth candidate model is misspecified.

Condition (C.3) $\widehat{\boldsymbol{V}} - \boldsymbol{V} = o_p(1)$.

Condition (C.4) *There exist vectors* $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_M^*$ *such that* $\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\| = O_p(d_m^{1/2} n^{-1/2})$, *for any* $m \in \{1, \ldots, M\}$ *and* $\max_{m \in \{1,\ldots,M\}} \|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\| = O_p(d^{1/2} M^{1/2} n^{-1/2})$, *where* $\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*\| = \{(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\}^{1/2}$.

Condition (C.5) *Uniformly for any* $\mathbf{w} \in \mathcal{W}$ *and any vector* $\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}$ *between* $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ *and* $\boldsymbol{\theta}^*(\mathbf{w})$,

$$\lambda_{\max}\left[\frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \Big|_{\widehat{\boldsymbol{\theta}}(\mathbf{w}) = \widetilde{\boldsymbol{\theta}}_{\mathbf{w}}}\right] = O_p(1),$$

where $\lambda_{\max}(\cdot)$ *denotes the largest singular value of a matrix.*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Condition (C.3) requires the estimator $\widehat{\boldsymbol{V}}$ to be consistent. Condition (C.4) is a high-level condition. When the candidate model m is correctly specified, the root-n consistency in Condition (C.4) has been shown by, for example, Harris & Mátyás (1999). When the candidate model m is misspecified and d_m is fixed, Hall & Inoue (2003) proved $\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^* = O_p(n^{-1/2})$, under some regularity conditions. Condition (C.5) requires that the largest singular value of the fixed-dimensional matrix $\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}/\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}|_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}}$ is uniformly bounded, and this matrix depends on the specific form of the working moment function $\boldsymbol{\mu}(\boldsymbol{\theta})$.

Let $\boldsymbol{\theta}^*(\mathbf{w}) = \sum_{m=1}^{M} w_m \boldsymbol{\Pi}_m \boldsymbol{\theta}_m^*$,

$$L^*(\mathbf{w}) = \left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right],$$

and $\xi_n = \inf_{\mathbf{w}\in\mathcal{W}} L^*(\mathbf{w})$.

Condition (C.6) $M^{1/2}p^{1/2}n^{-1/2}\xi_n^{-1} \to 0$.

Condition (C.6) requires that the minimum limitation loss decreases at a rate slower than $n^{-1/2}$ when $n \to \infty$. Similar conditions are used in Ando & Li (2014), Zhang et al. (2016), and Ando & Li (2017). To further discuss Condition (C.6), we first define a correctly specified model. For model $\widetilde{m}$, if there exists a value of $\boldsymbol{\theta}_{\widetilde{m}}$ such that $\boldsymbol{\mu}(\boldsymbol{\Pi}_{\widetilde{m}}^{\mathrm{T}}\boldsymbol{\theta}_{\widetilde{m}}) = \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)$, then we say that model $\widetilde{m}$ is correctly specified. If one of the candidate models (say

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

model $\widetilde{m}$) is correctly specified, then $\boldsymbol{\mu}(\boldsymbol{\Pi}_{\widetilde{m}}^{\mathrm{T}}\boldsymbol{\theta}_{\widetilde{m}}^{*}) = \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)$, and thus

$$L^{*}(\mathbf{w}_{\widetilde{m}}^{0}) = \left\{\boldsymbol{\mu}(\boldsymbol{\Pi}_{\widetilde{m}}^{\mathrm{T}}\boldsymbol{\theta}_{\widetilde{m}}^{*}) - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega} \left\{\boldsymbol{\mu}(\boldsymbol{\Pi}_{\widetilde{m}}^{\mathrm{T}}\boldsymbol{\theta}_{\widetilde{m}}^{*}) - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right] = 0, \quad (4.1)$$

where $\mathbf{w}_{\widetilde{m}}^{0}$ is an $M \times 1$ vector, in which the $\widetilde{m}th$ element is one and the others are zeros. Hence, Condition (C.6) requires that all candidate models are misspecified. This condition is commonly used to study the properties of an AIC-type model selection criterion; see, for example, Li (1987) and Shao (1997).

<u>Theorem</u> 1 *Under Conditions (C.1)–(C.6) and the conditions in Proposition 2, we have*

$$\frac{L(\widehat{\mathbf{w}})}{\inf\limits_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w})} \to 1 \quad (4.2)$$

in probability, where the squared loss function $L(\mathbf{w})$ is defined in (3.3).

The proof of Theorem 1 is provided in Section S.5 of the Supplementary Material. This theorem shows that the model averaging procedure using $\widehat{\mathbf{w}}$ is asymptotically optimal in the sense that the resulting squared loss is asymptotically identical to that of the infeasible best possible model average estimator.

4.2 Root-n consistency when there are correctly specified candidate models

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

The asymptotic optimality in Section 4.1 requires all candidate models are misspecified. However, in practice, we never know whether there are correctly specified candidate models (we say that the mth candidate model is correctly specified if there exists a value of $\boldsymbol{\theta}_m$ such that $\boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\boldsymbol{\theta}_m) = \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)$), which may happen. Hence, in this section, we provide theoretical support for our method when there are correctly specified candidate models. In this case, our method exhibits root-n consistency, which means that in a large-sample sense, our method at least does not perform worse than the commonly used methods that also achieve root-n consistency. We further impose the following regularity condition.

Condition (C.7) *Uniformly for any* $\mathbf{w} \in \mathcal{W}$ *and any vector* $\widetilde{\boldsymbol{\theta}}$ *between* $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ *and* $\boldsymbol{\theta}^*(\mathbf{w})$,

$$\lambda_{\min}^{-1}\left[\frac{\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})}^{\mathrm{T}}\,|_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}}\right] = O_p(1),$$

where $\lambda_{\min}(\cdot)$ *denotes the smallest singular value of a matrix.*

Condition (C.7) is similar to Condition (C.5), but requires that the smallest singular value of the matrix be bounded away from zero.

<u>Theorem</u> 2 *Under Conditions (C.1)–(C.5) and (C.7) and the conditions in Proposition 2, if there exists at least one correctly specified candidate*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

model (say model $\widetilde{m}$), then

$$\|\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\| = O_p(n^{-1/2}p^{1/2}). \tag{4.3}$$

The proof of Theorem 2 is provided in Section S.6 of the Supplementary Material. Combining Theorems 1−2, the proposed $\mathrm{MA}_{\mathrm{GMM}}$ method has a theoretical justification in a large-sample sense, regardless of whether or not there are correctly specified candidate models.

5. Monte Carlo

In this section, we conduct Monte Carlo experiments to examine the finite-sample performance of the proposed model averaging method based on the GMM ($\mathrm{MA}_{\mathrm{GMM}}$). Here, we compare the model selection estimator $\mathrm{MS}_{\mathrm{GMM}}$ and the GMM estimator. We do not compare our method with other existing selection or averaging methods because they focus on candidate models with different moment conditions. As stated in Section 2, the candidate models for our method use the same moment conditions, but different variables.

5.1 Data-generation process

We consider two simulation designs. In the first design, the true data-generation procedure is captured by at least one of the candidate models,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

while in the second, it is not; that is, all candidate models are misspecified in the second design.

Design I. We use the linear regression models with instrumental variables described in Remark 2. Specifically, we set

$$y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta} + \epsilon_i, \qquad \mathbf{X}_i = (1, Y_i, \mathbf{q}_i^{\mathrm{T}})^{\mathrm{T}},$$

$$\boldsymbol{\theta} = (1, 1, 0.2, -0.001, 1, 0.01, 0.2, 0.01)^{\mathrm{T}}, \qquad \mathbf{q}_i \sim \mathrm{Normal}\{\mathbf{0}_{6\times 1}, (0.5^{|j_1-j_2|})_{1\le j_1,j_2\le 6}\},$$

$$Y_i = \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\gamma} + u_i, \qquad \boldsymbol{\gamma} = \delta(1, 1, 1, 1, 1, 1, 1)^{\mathrm{T}},$$

$$\mathbf{h}_i \sim \mathrm{Normal}\{\mathbf{0}_{7\times 1}, (0.5^{|j_1-j_2|})_{1\le j_1,j_2\le 7}\}, \qquad \begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim \mathrm{Normal}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0.5\sigma \\ 0.5\sigma & 1 \end{pmatrix}\right\}.$$

Hence, the correlation coefficient between ϵ_i and u_i is 0.5, and the instrumental variable vector is $\mathbf{Z}_i = (1, \mathbf{h}_i^{\mathrm{T}}, \mathbf{q}_i^{\mathrm{T}})^{\mathrm{T}}$. We control σ^2 such that the theoretical $R^2 \equiv \mathrm{var}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})/\mathrm{var}(y_i)$ varies in the set $\{0.2, 0.3, \ldots, 0.8\}$, and control δ such that the theoretical $\widetilde{R}^2 \equiv \mathrm{var}(\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\gamma})/\mathrm{var}(Y_i)$ varies in the set $\{0.2, 0.5, 0.8\}$. The sample size n is set to 30, 80, 150, or 300. Here, we consider the case with a very small sample size, i.e., $n = 30$, because we find that when the sample size is large, all methods tend to perform very similarly. The variables in $\mathbf{q}_i$ are set to be auxiliary (i.e., they are possibly used in the candidate models); hence, we have $2^6 = 64$ candidate models.

To evaluate the methods, we use 10^4 replications. In each replication, we obtain the estimators of the coefficients of the endogenous variable Y_i

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

by using the GMM, $\mathrm{MA}_{\mathrm{GMM}}$, and $\mathrm{MS}_{\mathrm{GMM}}$, which is defined in the text following (3.9). As described in Remark 2, we set $\boldsymbol{\Omega} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n)^{-1}$ for all methods. Then, we calculate MSE using these 10^4 replications. To facilitate the comparisons, all MSEs are normalized using the MSE of the GMM.

Design II. In this design, we generate y_i as

$$y_i = X_{i1}\theta_1 + \cdots + X_{i6}\theta_6 + \exp(X_{i7}\theta_7) + \exp(X_{i8}\theta_8) + \epsilon_i,$$

where X_{ij} and θ_j are the jth components of $\mathbf{X}_i$ and $\boldsymbol{\theta}$, respectively. All other settings in Design II are the same as those in Design I. Hence, in this design, all candidate models are misspecified.

In contrast to Design I, we do not use the MSE in the coefficient estimation to evaluate the methods in this design, because the estimators may not all be consistent. Instead, we use the estimation loss, defined in (3.3), to evaluate the methods. Then, we calculate the mean loss using the 10^4 replications. To facilitate the comparisons, all losses are normalized using the loss from the GMM.

5.2 Results

The results of the simulations under Design I are presented in Figure 1 and Figures S.1−S.2 of the Supplementary Material. It is clear from the figures that when $n \in \{30, 80, 150\}$, the $\mathrm{MA}_{\mathrm{GMM}}$ yields the most accurate results for a very large range of values of R^2. When $n = 300$, the three

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

methods perform similarly, because there are correctly specified candidate models in this design. Thus, all three methods achieve root-n consistency. The $\mathrm{MS}_{\mathrm{GMM}}$ is always dominated by the $\mathrm{MA}_{\mathrm{GMM}}$. When $\widetilde{R}^2$ decreases, the three methods perform more disparately, and the R^2 range in which the $\mathrm{MA}_{\mathrm{GMM}}$ has an advantage over the GMM widens, compare the left-bottom panels of Figure 1 and Figure S.1 of the Supplementary Material.

The simulation results for Design II are presented in Figure 2 and Figures S.3−S.4 of the Supplementary Material. Again, we find that when R^2 is small or moderate, the $\mathrm{MA}_{\mathrm{GMM}}$ outperforms the GMM; when R^2 is large, the GMM can be superior to the $\mathrm{MA}_{\mathrm{GMM}}$. When the sample size is 300 and R^2 is close to 0.8, the $\mathrm{MS}_{\mathrm{GMM}}$ outperforms the $\mathrm{MA}_{\mathrm{GMM}}$. However, for all other settings, $\mathrm{MA}_{\mathrm{GMM}}$ performs best.

6. Empirical application

6.1 Data and models

We analyze data from the 1980 census on the median thousand dollar value of owner-occupied housing (hsngval) and the median monthly gross rent (rent) in the 50 US states. The data are provided by Stata: https://www.stata.com/. We model the rent as

$$rent_i = \theta_1 + \theta_2 hsngval_i + \theta_3 pcturban_i + \theta_4 region2_i + \theta_5 region3_i + \theta_6 region4_i + \epsilon_i, \quad (6.1)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

where "pcturban" is the percentage of the population living in urban areas, and "region2", "region3" and "region4" are dummy region variables. Because we focus on the impact of "hsngval" on "rent", we set the other variables ("pcturban", "region2", "region3" and "region4") to be auxiliary (i.e., they are possibly used in the candidate models). Hence, we have $2^4 = 16$ candidate models. Because we do not know whether all of these candidate models are misspecified, and our method has theoretical support regardless of whether this is the case, we use our method for this data set.

Because random shocks that affect rent in a state may also affect housing prices, the variable "hsngval" is taken as endogenous. The median of family income (faminc) and the region variables are used as instrumental variables; that is,

$$hsngval_i = \gamma_1 + \gamma_2 faminc_i + \gamma_3 region2_i + \gamma_4 region3_i + \gamma_5 region4_i + u_i. \quad (6.2)$$

Panel I of Table 1 shows the coefficient estimates of the main model (6.1). The effects estimated by the $\mathrm{MA}_{\mathrm{GMM}}$ are smaller than those of the GMM. The variables "region2" and "region3" are not selected by $\mathrm{MS}_{\mathrm{GMM}}$. Panel II of Table 1 shows the weights of the $\mathrm{MA}_{\mathrm{GMM}}$. The the weights are primarily assigned to four models, with the largest weight assigned to the model selected by the $\mathrm{MS}_{\mathrm{GMM}}$.

6.2 Comparison of estimation performance

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Table 1: Coefficient estimates and weights in the real-data analysis. The notation * indicates that the model includes the corresponding variable. For example, Model 1 only includes "constant" and "hsngval".

	Panel I			Panel II			
	Coefficient estimates			Weights of models with weights larger than 10^{-4}			
Variables	GMM	$\mathrm{MS}_{\mathrm{GMM}}$	$\mathrm{MA}_{\mathrm{GMM}}$	Model 1	Model 2	Model 3	Model 4
constant	88.3141	96.7447	94.7084	*	*	*	*
hsngval	3.8691	3.7037	3.5430	*	*	*	*
pcturban	-0.4993	-0.4612	-0.3414			*	*
region2	1.5253	-	0.0000				
region3	7.7394	-	2.1899				*
region4	-40.6289	-41.0891	-36.8204		*	*	*
Weights				0.0586	0.2247	0.3931	0.3235

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

To compare the three methods using the real data, we generate data by sampling the residuals. Specifically, let $\widehat{\boldsymbol{\gamma}}_{\text{OLS}}$ be the ordinary least squares estimator of coefficients in model (6.2). The residual is

$$\widehat{u}_i = hsngval_i - (1, faminc_i, region2_i, region3_i, region4_i)\widehat{\boldsymbol{\gamma}}_{\text{OLS}}, \qquad (6.3)$$

for $i = 1, \ldots, 50$. By sampling in $\{\widehat{u}_1, \ldots, \widehat{u}_{50}\}$ 50 times with repetition, we obtain $\widehat{u}_1^{(r)}, \ldots, \widehat{u}_{50}^{(r)}$. Then, we obtain

$$hsngval_i^{(r)} = (1, faminc_i, region2_i, region3_i, region4_i)\widehat{\boldsymbol{\gamma}}_{\text{OLS}} + \widehat{u}_i^{(r)}.$$

Let $\widehat{\boldsymbol{\theta}}_{\text{Method}}$ be the estimator of the coefficients in model (6.1), where Method is GMM, MS_{GMM}, or MA_{GMM}. The estimators are shown in Panel I of Table 1. Similarly to (6.3), we obtain the residual

$$\widehat{\epsilon}_i = rent_i - (1, hsngval_i, pcturban_i, region2_i, region3_i, region4_i)\widehat{\boldsymbol{\theta}}_{\text{Method}},$$

for $i = 1, \ldots, 50$. By sampling in $\{\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_{50}\}$ 50 times with repetition, we obtain $\widehat{\epsilon}_1^{(r)}, \ldots, \widehat{\epsilon}_{50}^{(r)}$. Then the response variable in the main model is generated by

$$rent_i^{(r)} = (1, hsngval_i^{(r)}, pcturban_i, region2_i, region3_i, region4_i)\widehat{\boldsymbol{\theta}}_{\text{Method}} + \widehat{\epsilon}_i^{(r)}.$$

We generate 10^4 data sets; that is $r = 1, \ldots, 10^4$. Table 2 shows the MSE when estimating the coefficient of the endogenous variable "hsngval" based

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

on the 10^4 replications. Regardless of which estimated coefficients are used to generate the data sets, the proposed $\mathrm{MA}_{\mathrm{GMM}}$ method always performs best.

Lastly, we compare the out-of-sample prediction performance of the different methods. We randomly divide the 50 observations into a training sample of n_1 observations and a test sample of $n - n_1$ observations. We set $n_1 \in \{20, 30, 40\}$. The predictions of the three methods are based on model (6.1). The average squared prediction errors are calculated across observations in the test sample. We randomly divide the sample into training and test samples 10^4 times. Table 3 provides the mean of the average squared prediction errors based on the 10^4 replications. Regardless of how the sample is divided, the proposed $\mathrm{MA}_{\mathrm{GMM}}$ method always performs best.

Table 2: MSE in estimating the coefficient of the endogenous variable "hsngval".

	GMM	$\mathrm{MS}_{\mathrm{GMM}}$	$\mathrm{MA}_{\mathrm{GMM}}$
$\widehat{\boldsymbol{\theta}}_{\mathrm{Method}}$ is from GMM	0.7056	0.7101	0.6289
$\widehat{\boldsymbol{\theta}}_{\mathrm{Method}}$ is from $\mathrm{MS}_{\mathrm{GMM}}$	0.6486	0.6465	0.5660
$\widehat{\boldsymbol{\theta}}_{\mathrm{Method}}$ is from $\mathrm{MA}_{\mathrm{GMM}}$	0.5977	0.5987	0.5192

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Table 3: Mean of average squared prediction errors $\times 10^{-4}$.

	GMM	$\mathrm{MS}_{\mathrm{GMM}}$	$\mathrm{MA}_{\mathrm{GMM}}$
$n_1 = 20$	2.6015	0.5914	0.5482
$n_1 = 30$	0.1845	0.1529	0.1400
$n_1 = 40$	0.1462	0.1375	0.1249

7. Conclusion

In this paper, we propose optimal model averaging based on the GMM. Theoretical justifications are provided, regardless of whether all of the candidate models are misspecified. The numerical examples also show the promise of the proposed method. While the results in this paper offer some interesting insights to the application of the GMM, they also raise some important issues that warrant further study.

First, in general, under the GMM framework, the candidate models can vary with respect to (1) the moment restrictions and (2) the specification of a working moment function. In this study, we ignore the first situation. The proposed weight choice method cannot be used in this situation because our method depends heavily on the loss function (3.3). If the moment restrictions vary with working models, then the true moment $\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)$ in

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

(3.3) can do so as well, leading to serious difficulty in defining a reasonable loss function. Developing an asymptotically optimal model averaging method under this situation warrants future study.

Second, when there are correctly specified candidate models, we only derive the root-n convergence rate for the true parameter vector. We cannot establish its limit distribution theory, owing to the difficulties caused by the random weights. Hjort & Claeskens (2003) and Zhang & Liu (2019) may serve as useful guides in this regard. However, studies that follow Hjort & Claeskens (2003) use the locally misspecified moment conditions; see DiTraglia (2016), for example. In Zhang & Liu (2019), the nested setup of the candidate models limits the flexibility of their theory. Much future effort is required to promote research on inferences after averaging GMM estimators.

Lastly, in this study, the dimension of $\widehat{\boldsymbol{\mu}}$ is fixed. When the dimension of $\widehat{\boldsymbol{\mu}}$ is divergent to infinity with n, Proposition 2 still holds. For Proposition 1, we conjecture that the criterion $\widetilde{C}(\mathbf{w})$ is still an approximately unbiased estimator of the risk, although this requires a more detailed derivation. Additionally, we think that the optimality and consistency can be derived using techniques similar to those in the current proofs. Detailed derivations warrant future study.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

8. Acknowledgments

The author is grateful to the co-Editor Hans-Georg Müller, and to the associate editor and two referees for their constructive comments. Zhang's research was partially supported by the National Natural Science Foundation of China (Grant nos. 71925007, 71631008, and 11688101), the Youth Innovation Promotion Association of Chinese Academy of Sciences, the Beijing Academy of Artificial Intelligence, and a joint grant from the Academy for Multidisciplinary Studies, Capital Normal University. This work occurred when the author visited Penn State University.

9. Supplementary Material

The online Supplementary Material contains the technical proofs and figures for the outcomes of the simulation studies.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

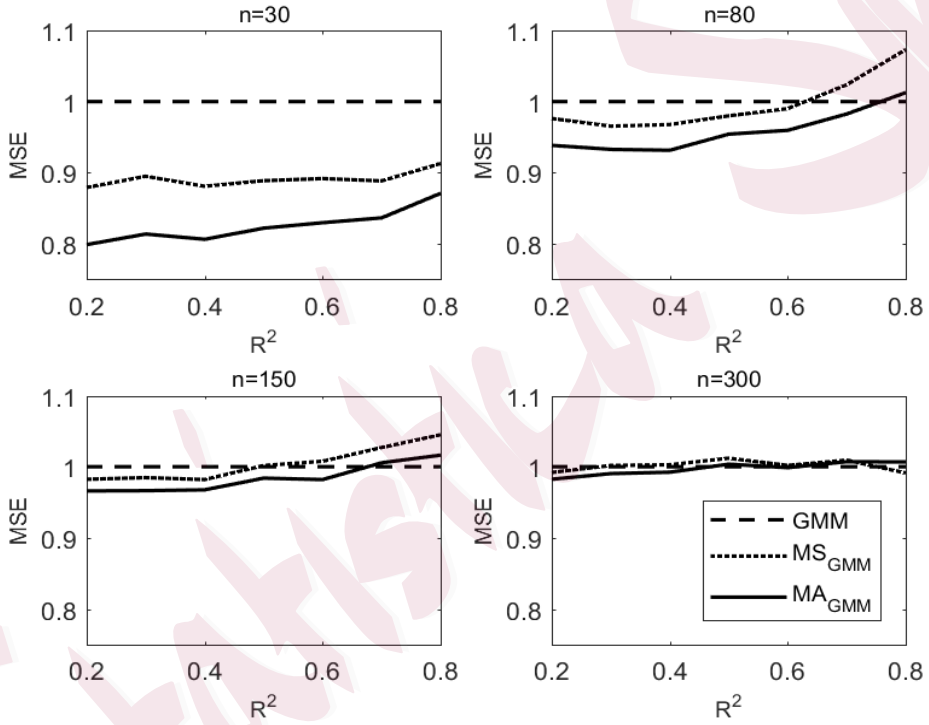

Figure 1: MSE in simulation Design I, with $\widetilde{R}^2 = 0.2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

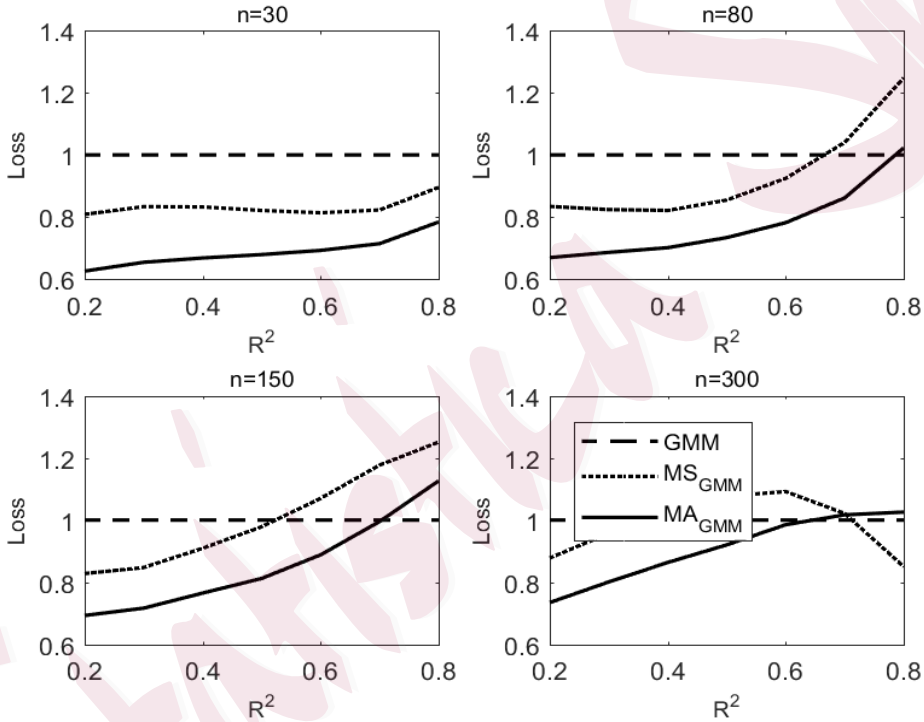

Figure 2: Loss in simulation Design II, with $\widetilde{R}^2 = 0.2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

References

Ando, T. & Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* **109**, 254–265.

Ando, T. & Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics* **45**, 2654–2679.

Andrews, D. W. K. (1991). Asymptotic optimality of generalized c_l, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics* **47**, 359–377.

Bates, J. M. & Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly* **20**, 451–468.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

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

Chen, J., Li, D., Linton, O. & Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association* **113**, 919–932.

Cheng, X. & Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics* **186**, 280–293.

Cheng, X., Liao, Z. & Shi, R. (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics* **10**, 931–979.

De Luca, G., Magnus, J. R. & Peracchi, F. (2018). Weighted-average least squares estimation of generalized linear models. *Journal of Econometrics* **204**, 1–17.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Den Haan, W. J. & Levin, A. T. (1997). A practitioner's guide to robust covariance matrix estimation. *Handbook of statistics* **15**, 299–342.

DiTraglia, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for gmm. *Journal of Econometrics* **195**, 187–208.

Fang, F., Lan, W., Tong, J. & Shao, J. (2019). Model averaging for prediction with fragmentary data. *Journal of Business & Economic Statistics* **37**, 517–527.

Hall, A. R. & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* **114**, 361–394.

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

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* **5**, 495–530.

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

Harris, D. & Mátyás, L. (1999). *Introduction to the Generalized Method of Moments Estimation*. Themes in Modern Econometrics. Cambridge University Press.

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

Hjort, N. L. & Claeskens, G. (2006). Focused information criteria and model averaging for the cox hazard regression model. *Journal of the American Statistical Association* **101**, 1449–1464.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Leung, G. & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52**, 3396–3410.

Li, D., Linton, O. & Lu, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187**, 345–357.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics* **15**, 958–975.

Liang, H., Zou, G., Wan, A. T. K. & 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* **186**, 142–159.

Liu, Q. & Okui, R. (2013). Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal* **16**, 463–472.

Liu, Q., Okui, R. & Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews* **35**, 1692–1752.

Longford, N. T. (2005). Editorial: Model selection and efficiency-is 'which model?' the right question? *Journal of the Royal Statistical Society, Series A* **168**, 469–472.

Magnus, J. R., Powell, O. & Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* **154**, 139–153.

Magnus, J. R., Wan, A. T. K. & Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics & Data Analysis* **55**, 1331–1341.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0230

Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

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

Wan, A. T. K., Zhang, X. & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* **156**, 277–283.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–588.

Yin, S.-Y., Liu, C.-A. & Lin, C.-C. (2019). Focused information criterion and model averaging for large panels with a multifactor error structure. *Journal of Business & Economic Statistics* , 1–44.

Yuan, Z. & Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* **100**, 1202–1214.

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

Zhang, X. & Liu, C.-A. (2019). Inference after model averaging in linear regression models. *Econometric Theory* **35**, 816–841.

Zhang, X., Lu, Z. & Zou, G. (2013). Adaptively combined forecasting for discrete response time series. *Journal of Econometrics* **176**, 80–91.

Zhang, X. & Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica* **29**, 693–718.

Zhang, X., Yu, D., Zou, G. & Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* **111**, 1775–1790.

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