

Article

Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging

Naoya Sueishi

Graduate School of Economics, Kyoto University, Yoshida-Hommachi, Sakyo-ku, Kyoto, 6068501, Japan; E-Mail: sueishi@econ.kyoto-u.ac.jp; Tel.: +81-75-753-3500

Received: 13 May 2013; in revised form: 26 June 2013 / Accepted: 27 June 2013 /

Published: 3 July 2013

Abstract: This paper develops model selection and averaging methods for moment restriction models. We first propose a focused information criterion based on the generalized empirical likelihood estimator. We address the issue of selecting an optimal model, rather than a correct model, for estimating a specific parameter of interest. Then, this study investigates a generalized empirical likelihood-based model averaging estimator that minimizes the asymptotic mean squared error. A simulation study suggests that our averaging estimator can be a useful alternative to existing post-selection estimators.

Keywords: model selection; model averaging; focused information criterion; generalized empirical likelihood

1. Introduction

This paper develops model selection and averaging methods for moment restriction models. We first propose a focused information criterion (FIC) based on the generalized empirical likelihood (GEL) estimator [1,2], which nests the empirical likelihood (EL) [3,4] and exponential tilting (ET) [5,6] estimators as special cases. Motivated by Claeskens and Hjort [7], we address the issue of selecting an optimal model for estimating a specific parameter of interest, rather than identifying a correct model or selecting a model with good global fit. Then, as an extension of FIC, this study presents a GEL-based frequentist model averaging (FMA) estimator that is designed to minimize the mean squared error (MSE) of the estimator.

Traditional model selection methods, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), select a single model regardless of the specific goal of inference [8,9]. AIC

selects a model that is close to the true data generating process (DGP) in terms of Kullback-Leibler discrepancy, while BIC selects the model with the highest posterior probability. However, a model with good global fit is not necessarily a good model for estimating a specific parameter. For instance, Hansen [10] considers the problem of deciding the order of autoregressive models. His simulation study demonstrates that the AIC-selected model does not necessarily produce a good estimate of the impulse response. This result reveals that the best model generally differs for different intended uses of the model.

In their seminal work, Claeskens and Hjort [7] established an FIC that is designed to select the optimal model depending on its intended use. Their goal is to select the model that attains the minimum MSE of the maximum likelihood estimator for the parameter of interest, which they call the focus parameter. The FIC is constructed from an asymptotic estimate of the MSE.

Since then, an FIC has been derived for several models. Claeskens, Croux and Kerckhoven [11] proposed an FIC for logistic regressions. Hjort and Claeskens [12] proposed an FIC for the Cox hazard regression model. Zhang and Liang [13] developed an FIC for the generalized additive partial linear model. Models studied in those papers are likelihood-based. However, econometric models are often specified via moment restrictions rather than parametric density functions. This paper indicates that the idea of Claeskens and Hjort [7] is applicable to moment restriction models. Our FIC is constructed using an asymptotic estimate of the MSE of the GEL estimator.

Model selection for moment restriction models is still underdeveloped. Andrews and Lu [14] proposed selection criteria based on the J -statistic of the generalized method of moments (GMM) estimator [15]. Hong, Preston and Shum [16] extended the results of Andrews and Lu to the GEL estimation. Sueishi [17] developed information criteria similar to the AIC. The goal of Andrews and Lu [14] and Hong, Preston and Shum [16] was to identify the correct model, whereas Sueishi [17] selects the best approximating model in terms of Cressie-Read discrepancy. Although these criteria are useful in many applications, they do not address the issue of selecting the model that best serves its intended purpose.

Model averaging is an alternative to model selection. Inference after model selection is typically conducted as if the selected model is the true DGP. However, this ignores uncertainty introduced by model selection. Rather than conditioning on the single selected model, the averaging technique uses all candidate models to incorporate model selection uncertainty. Although Bayesian methods are predominant in the literature [18], there is also a growing FMA literature for likelihood-based models [19–21]. See also Yang [22], Leung and Barron [23] and Goldenshluger [24] for related issues.

In the FMA literature, it is often of particular interest to obtain an optimal averaging estimator in terms of a certain loss [25–28]. This study investigates a GEL-based averaging method that minimizes the asymptotic mean squared error in a framework similar to that of Hjort and Claeskens [21]. A simulation study indicates that our averaging estimator outperforms existing post-model-selection estimators.

Although this study investigates GEL-based methods, in general, its results are readily applied to the two-step GMM estimator, because our results rely only on first-order asymptotic theory. However, the two-step GMM estimator often suffers from a large bias that cannot be captured by first-order asymptotics, even if the model is correctly specified. Because the FIC addresses a trade-off between

misspecification bias and estimation variance, the GEL estimator will be more suitable for our framework.

Now, we review related works. DiTraglia [29] proposes an instrument selection criterion for GMM that is based on the concept of FIC. Our approach resembles DiTraglia's, but his interest is instrument selection, whereas ours is model selection. DiTraglia intentionally uses an invalid large set of instruments to improve efficiency; we intentionally use a wrong small model to improve efficiency. Liu [30] proposes an averaging estimator for the linear regression model by using a local asymptotic framework. Although Liu considers exogenous regressors, we allow endogenous regressors. Martins and Gabriel [31] consider GMM-based model averaging estimators under a framework different from ours.

The remainder of the paper is organized as follows. Section 2 describes our local misspecification framework. Section 3 derives the FIC. Section 4 discusses the FMA estimator. Section 5 provides a simple example for which our methods are applicable. Section 6 presents the result of Monte Carlo study. Section 7 concludes.

2. Local Misspecification Framework

We first introduce our setup. The basic construct follows Claeskens and Hjort [7]. There is a smallest and a largest model in our set of candidate models. The smallest, which we call the reduced model, has a p dimensional unknown parameter vector, $\theta = (\theta_1, \dots, \theta_p)'$. The largest, or the full model, has an additional q dimensional unknown parameter vector, $\gamma = (\gamma_1, \dots, \gamma_q)'$. The full model is assumed to be correctly specified and nests the reduced model; *i.e.*, the reduced model corresponds to the special case of the full model in which $\gamma = \gamma_0 = (\gamma_{0,1}, \dots, \gamma_{0,q})'$ for some known γ_0 . Typically, γ_0 is a vector of zeros: $\gamma_0 = (0, \dots, 0)'$. An example is given in Section 5.

There are up to 2^q submodels, all of which have θ as the common parameter vector. A submodel treats some elements of γ as unknown parameters and is indexed by a subset, S , of $\{1, \dots, q\}$. The model, S , contains parameters, γ_j , such that $j \in S$. Thus, the reduced and full models correspond to $S = \emptyset$ and $S = \{1, \dots, q\}$, respectively. We use “red” and “full” to denote the reduced and full models, respectively.

The focus parameter, μ , which is the parameter of interest, is a function of θ and γ : $\mu = \mu(\theta, \gamma)$. It could be merely an element of θ . Prior knowledge or economic theories suggest that θ should be estimated, but we are unsure which elements of γ should be treated as unknown parameters. Estimating a larger model usually implies a lesser modeling bias and a larger estimation variance. However, if the reduced model is globally misspecified in the sense that the violation of the moment restriction does not disappear even in the limit, then the misspecification bias asymptotically dominates the variance of the GEL estimator. Thus, we cannot make a reasonable comparison of bias and variance in the asymptotic framework.

A local misspecification framework is introduced to take into account the bias-variance trade-off. Let $y_1, \dots, y_n$ be i.i.d. random vectors from an unknown density, $f_n(y)$, which depends on the sample size,

n .¹The functional form of $f_n(y)$ is not specified. The full model is defined via the following moment restriction:

$$E_n [m(y_i, \theta_0, \gamma_0 + \delta/\sqrt{n})] \equiv \int m(y, \theta_0, \gamma_0 + \delta/\sqrt{n}) f_n(y) dy = 0, \quad (1)$$

where $m : \mathbb{R}^{d_y} \times \Theta \times \Gamma \rightarrow \mathbb{R}^l$ is a known vector-valued function up to the parameters. For each n , the true parameter values of θ and γ are θ_0 and $\gamma_0 + \delta/\sqrt{n}$, respectively. Note that γ_0 is known, but θ_0 and δ are unknown. We assume that $l > p + q$; i.e., the model is over-identified.

The moment function of the reduced model is $m(y, \theta, \gamma_0)$. The reduced model is misspecified in the sense that there is no value $\theta^* \in \Theta$, such as $E_n[m(y_i, \theta^*, \gamma_0)] = 0$, for any fixed n . However, if the moment function is differentiable with respect to γ , then (1) implies that the reduced model satisfies:

$$\|E_n [m(y_i, \theta_0, \gamma_0)]\| = \left\| E_n \left[\frac{\partial m(y_i, \theta_0, \bar{\gamma})}{\partial \gamma'} \right] \delta / \sqrt{n} \right\| = O(1/\sqrt{n})$$

for some vector, $\bar{\gamma}$ between γ_0 and $\gamma_0 + \delta/\sqrt{n}$. Thus, even though the moment restriction is invalid at (θ_0, γ_0) , the violation disappears in the limit. A similar relationship also holds for the other submodels. As the next section reveals, under this framework, the squared bias and variance of the GEL estimator are both of the order, $O(1/n)$. Hence, the trade-off between bias and variance can be considered. If δ is sufficiently small, it might be better to set $\gamma = \gamma_0$ rather than estimate γ .

In general, the dimension of the moment function can differ among submodels. For instance, consider a linear instrumental variable model. The model (structural form) can be estimated as long as the number of instruments exceeds or equals the number of unknown parameters. Thus, it is possible to use only a subset of instruments to estimate a submodel. For ease of exposition, however, we consider only the case where the dimension of the moment function is fixed for all submodels.

3. Focused Information Criterion

To construct an FIC, we first derive the asymptotic distribution of the GEL estimator under the local misspecification framework. Newey [32] and Hall [33] obtained a similar result in the case of GMM estimation to analyze the local power properties of specification tests.

A model, S , contains $p + q_S$ unknown parameters. The moment function of the model is denoted as $m(y, \theta, \gamma_S) = m(y, \theta, \gamma_S, \gamma_{0,S^C})$, where S^C is the complementary set of S . The values of γ_j are set to be their null values $\gamma_{0,j}$ for $j \in S^C$.

Let $\rho(v)$ be a concave function on its domain, $\mathcal{V}$, which is an open interval containing zero. We normalize $\rho(v)$, so that $\rho_1(0) = \rho_2(0) = -1$, where $\rho_j(v) = d^j \rho(v) / dv^j$. The GEL estimator of (θ, γ_S) is obtained by solving the saddle-point problem:

$$(\hat{\theta}_S, \hat{\gamma}_S) = \arg \min_{\theta \in \Theta, \gamma_S \in \Gamma_S} \max_{\tau \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \rho(\tau' m(y_i, \theta, \gamma_S)),$$

where $\Gamma_S \subset \mathbb{R}^{q_S}$ is the parameter space of γ_S and $\mathcal{T} \subset \mathbb{R}^l$ is the set of feasible values of τ . The EL and ET estimators are special cases with $\rho(v) = \log(1 - v)$ and $\rho(v) = -\exp(v)$, respectively. Although $\hat{\theta}_S$

¹Although $y_1, \dots, y_n$ is a triangular array, we suppress the additional subscript, n , on y for notational simplicity.

has p elements for any S , we adopt the subscript, S , to emphasize that the value of the estimator depends on S .

Let $m_i = m(y_i, \theta_0, \gamma_0)$, $m_{\theta i} = \frac{\partial m(y_i, \theta, \gamma)}{\partial \theta'} \Big|_{\theta=\theta_0, \gamma=\gamma_0}$, and $m_{\gamma i} = \frac{\partial m(y_i, \theta, \gamma)}{\partial \gamma'} \Big|_{\theta=\theta_0, \gamma=\gamma_0}$. Furthermore, let $m_{\gamma Si} = \frac{\partial m(y_i, \theta, \gamma)}{\partial \gamma'_S} \Big|_{\theta=\theta_0, \gamma=\gamma_0}$. We define:

$$J_S = \begin{pmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix} = \begin{pmatrix} E[m_{\theta i}]' E[m_i m_i']^{-1} E[m_{\theta i}] & E[m_{\theta i}]' E[m_i m_i']^{-1} E[m_{\gamma Si}] \\ E[m_{\gamma Si}]' E[m_i m_i']^{-1} E[m_{\theta i}] & E[m_{\gamma Si}]' E[m_i m_i']^{-1} E[m_{\gamma Si}] \end{pmatrix},$$

where E denotes the expectation with respect to $f(y) \equiv \lim_{n \rightarrow \infty} f_n(y)$. It is assumed that $f(y)$ satisfies:

$$E[m_i] = \int m(y, \theta_0, \gamma_0) f(y) dy = 0.$$

For the full model, we denote:

$$J_{\text{full}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}.$$

Then, we can write $J_{01,S} = J_{01} \pi'_S$ and $J_{11,S} = \pi_S J_{11} \pi'_S$, where π_S is the projection matrix of size, $q_S \times q$, that maps γ to the subvector, γ_S : $\pi_S \gamma = \gamma_S$.

Let $\hat{Q}_n(\theta, \gamma, \tau) = n^{-1} \sum_{i=1}^n \rho(\tau' m(y_i, \theta, \gamma))$ and $Q_n(\theta, \gamma, \tau) = E_n[\rho(\tau' m(y_i, \theta, \gamma))]$. Furthermore, let $Q(\theta, \gamma, \tau) = E[\rho(\tau' m(y_i, \theta, \gamma))]$. To obtain the asymptotic distribution of the GEL estimator, we impose the following conditions:

Assumption 3.1

1. $\Theta \subset \mathbb{R}^p$, $\Gamma \subset \mathbb{R}^q$, and $\mathcal{T} \subset \mathbb{R}^l$ are compact.
2. $m(y, \theta, \gamma)$ is continuous in $\theta \in \Theta$ and $\gamma \in \Gamma$ for almost every y .
3. $\sup_{\theta \in \Theta, \gamma \in \Gamma, \tau \in \mathcal{T}} |\hat{Q}_n(\theta, \gamma, \tau) - Q_n(\theta, \gamma, \tau)| \xrightarrow{p} 0$ under the sequence of $f_n(y)$.
4. $|Q_n(\theta, \gamma, \tau) - Q(\theta, \gamma, \tau)| \rightarrow 0$ as $n \rightarrow \infty$ for all $\theta \in \Theta$, $\gamma \in \Gamma$, and $\tau \in \mathcal{T}$.
5. $E[m(y_i, \theta, \gamma) m(y_i, \theta, \gamma)']$ is nonsingular for all $\theta \in \Theta$ and $\gamma \in \Gamma$.
6. (θ_0, γ_0) is the unique solution to $E[m(y_i, \theta, \gamma)] = 0$ and $(\theta_0, \gamma_0) \in \text{int}(\Theta \times \Gamma)$.
7. $\rho(v)$ is twice continuously differentiable in a neighborhood of zero.
8. $E[m_{\theta i}]$ and $E[m_{\gamma i}]$ are of full rank.
9. $\sup_n E_n[\|m_i\|^{2+\alpha}] < \infty$ for some $\alpha > 0$.
10. $m(y, \theta, \gamma)$ is continuously differentiable in θ and γ in a neighborhood, $\mathcal{N}$, of (θ_0, γ_0) .
11. $\sup_{\theta, \gamma \in \mathcal{N}} \left| n^{-1} \sum_{i=1}^n \frac{\partial m(y_i, \theta, \gamma)}{\partial \theta'} - E_n \left[\frac{\partial m(y_i, \theta, \gamma)}{\partial \theta'} \right] \right| \xrightarrow{p} 0$ and $\sup_{\theta, \gamma \in \mathcal{N}} \left| n^{-1} \sum_{i=1}^n \frac{\partial m(y_i, \theta, \gamma)}{\partial \gamma'} - E_n \left[\frac{\partial m(y_i, \theta, \gamma)}{\partial \gamma'} \right] \right| \xrightarrow{p} 0$ under the sequence of $f_n(y)$.
12. $\|E_n[m_{\theta i}] - E[m_{\theta i}]\| \rightarrow 0$ and $\|E_n[m_{\gamma i}] - E[m_{\gamma i}]\| \rightarrow 0$ as $n \rightarrow \infty$.

13. $\|E_n[m_i m_i'] - E[m_i m_i']\| \rightarrow 0$ as $n \rightarrow \infty$.

Conditions are rather high-level and strong. Some conditions can be replaced with primitive and weaker conditions [34].

We obtain the following lemma.

Lemma 3.1 Suppose Assumption 3.1 holds. Then, under the sequence of $f_n(y)$, we have:

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_S - \theta_0 \\ \hat{\gamma}_S - \gamma_{0,S} \end{pmatrix} \xrightarrow{d} N \left(J_S^{-1} \begin{pmatrix} J_{01} \\ \pi_S J_{11} \end{pmatrix} \delta, J_S^{-1} \right).$$

The proof is given in the Appendix.

If the model, S , is correctly specified, then the limiting distribution of the GEL estimator is $N(0, J_S^{-1})$. Therefore, as usual, local misspecification affects only the mean of the limiting distribution.

Next, we get the asymptotic distribution of the GEL estimator for the focus parameter. Additional notations are introduced. Let $Q = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$ and $Q_S = (\pi_S Q^{-1} \pi_S')^{-1}$; i.e., Q and Q_S are the lower right block matrices of J_{full}^{-1} and J_S^{-1} , respectively. Let $G_S = \pi_S' Q_S \pi_S Q^{-1}$. We assume that $\mu(\theta, \gamma)$ is differentiable with respect to θ and γ . Let:

$$w = J_{10}J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma} \quad \text{and} \quad \tau_0^2 = \left(\frac{\partial \mu}{\partial \theta} \right)' J_{00}^{-1} \frac{\partial \mu}{\partial \theta},$$

where the partial derivatives are evaluated at (θ_0, γ_0) . The true focus parameter is denoted as $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. Moreover, the GEL estimator of μ for the model, S , is denoted as $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$. Lemma 3.1 and the delta method imply the following theorem:

Theorem 3.1 Suppose Assumption 3.1 holds. Then, under the sequence of $f_n(y)$, we have:

$$D_n \equiv \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D \sim N(\delta, Q)$$

and:

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S \equiv \Lambda_0 + w'(\delta - G_S D),$$

where $\Lambda_0 \sim N(0, \tau_0^2)$ is independent of D .

The proof is almost the same as that of Lemma 3.3 in Hjort and Claeskens [21], so it is omitted.

Because $G_{\text{full}} = I$ and $G_{\text{red}} = 0$, as the special cases of the theorem, we have:

$$\begin{aligned} \sqrt{n}(\hat{\mu}_{\text{full}} - \mu_{\text{true}}) &\xrightarrow{d} N(0, \tau_0^2 + w'Qw), \\ \sqrt{n}(\hat{\mu}_{\text{red}} - \mu_{\text{true}}) &\xrightarrow{d} N(w'\delta, \tau_0^2). \end{aligned}$$

Therefore, in terms of the asymptotic MSE, the reduced model is better than the full model if $(w'\delta)^2 < w'Qw$, which is the case when the deviation of the reduced model from the true DGP is small.

More generally, Theorem 3.1 implies that the MSE of the limiting distribution of $\hat{\mu}_S$ is:

$$\text{mse}(S, \delta) = \tau_0^2 + w'\pi_S' Q_S \pi_S w + w'(I - G_S)\delta\delta'(I - G_S)'w. \quad (2)$$

The idea behind FIC is to estimate (2) for each model and select the model that attains the minimum estimated MSE.

All components in (2) except δ can be estimated easily by using their sample analogs. However, a consistent estimator for δ is unavailable, because D_n converges in distribution to a normal random variable. This difficulty is inevitable, as long as we utilize the local misspecification framework. Because the mean of DD' is $\delta\delta' + Q$, following Claeskens and Hjort [7], we use $D_n D_n' - \hat{Q}$ to estimate $\delta\delta'$. Then, the sample counterpart of (2) is:

$$\begin{aligned}\widehat{\text{mse}}(S) &= \hat{\tau}_0^2 + \hat{w}'\pi_S\hat{Q}_S\pi_S\hat{w} + \hat{w}'(I - \hat{G}_S)(D_n D_n' - \hat{Q})(I - G_S)'\hat{w} \\ &= \hat{w}'(I - \hat{G}_S)D_n D_n'(I - \hat{G}_S)'\hat{w} + 2\hat{w}'\pi_S'\hat{Q}_S\pi_S\hat{w} + \hat{\tau}_0^2 - \hat{w}'\hat{Q}\hat{w},\end{aligned}$$

which is an asymptotically unbiased estimator for (2). Because the last two terms do not depend on the model, we can ignore them for the purpose of model selection. Let $\hat{\psi}_{\text{full}} = \hat{w}'D_n$ and $\hat{\psi}_S = \hat{w}'\hat{G}_S D_n$. Then, our FIC for the model, S , is:

$$\begin{aligned}\text{FIC}_S &= \hat{w}'(I - \hat{G}_S)D_n D_n'(I - \hat{G}_S)'\hat{w} + 2\hat{w}'\pi_S'\hat{Q}_S\pi_S\hat{w} \\ &= (\hat{\psi}_{\text{full}} - \hat{\psi}_S)^2 + 2\hat{w}'\hat{Q}_S\hat{w}_S,\end{aligned}\quad (3)$$

where $\hat{w}_S = \pi_S\hat{w}$. The bigger the model is, the smaller the first term and the larger the second term in (3). Since w depends on μ , FIC can be used to select an appropriate submodel, depending on the parameter of interest.

Although we consider only the case where μ is a scalar, our FIC is also applicable to a vector-valued focus parameter by viewing each element of the vector as a different scalar-valued focus parameter. Different models might be used to estimate different elements of the vector.

We conclude this section with a remark on the estimation of $\delta\delta'$. Because we estimate $\delta\delta'$ by $D_n D_n' - \hat{Q}$, the estimate can be negative definite in finite sample. That means that the squared bias term can be negative. To avoid such cases, as suggested by Claeskens and Hjort [35], we can also use the following bias-corrected FIC:

$$\text{FIC}_S^* = \begin{cases} \text{FIC}_S & \text{if } N_n(S) \text{ does not take place} \\ \hat{w}'(I + \hat{G}_S)\hat{Q}\hat{w} & \text{if } N_n(S) \text{ takes place,} \end{cases}$$

where $N_n(S)$ is the event of negligible bias:

$$\left\{ \hat{w}'(I - \hat{G}_S)\hat{\delta}_{\text{full}} \right\}^2 < \hat{w}' \left(\hat{Q} - \pi_S'\hat{Q}_S\pi_S \right) \hat{w}.$$

See Section 6.4 of Claeskens and Hjort [35] for details.

4. Model Averaging

This section extends the result of Section 3 to the averaging problem. In the FMA literature, it is often of particular interest to obtain an optimal averaging estimator in terms of a certain loss. We consider a possibility of obtaining the best averaging weights that minimize the MSE in the local misspecification framework. A similar analysis is presented in Liu [30] in the case of linear regression.

Let $\mathcal{A}$ be the set of all candidate models. We consider an averaging estimator for the focus parameter of the form:

$$\hat{\mu} = \sum_{S \in \mathcal{A}} c(S) \hat{\mu}_S,$$

where the weights, $c(S)$, add up to unity. Note that a post-selection estimator of μ can also be written in this form. Let S_{FIC} be the FIC-selected model. Then the post-selection estimator using FIC is:

$$\hat{\mu}_{\text{FIC}} = \sum_{S \in \mathcal{A}} 1(S = S_{\text{FIC}}) \hat{\mu}_S,$$

where $1(\cdot)$ is the indicator function. Thus, the post-selection estimator is a special case of the averaging estimator.

If the weights are not random, then it is straightforward from Theorem 3.1 that:

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \xrightarrow{d} \Lambda \equiv \sum_{S \in \mathcal{A}} c(S) \Lambda_S \stackrel{d}{=} \Lambda_0 + w'(\delta - \hat{\delta}(D)),$$

where $\hat{\delta}(D) = \sum_{S \in \mathcal{A}} c(S) G_S D$. Therefore, the asymptotic mean and variance of the averaging estimator are given by:

$$\begin{aligned} E[\Lambda] &= \sum_{S \in \mathcal{A}} c(S) w'(I - G_S) \delta, \\ \text{Var}[\Lambda] &= \tau_0^2 + \sum_{S, S' \in \mathcal{A}} c(S) c(S') w'(G_S Q G_{S'}') w. \end{aligned}$$

Thus, there is a set of weights that minimizes the asymptotic MSE of $\hat{\mu}$.

Suppose there are M candidate models: $S_1, \dots, S_M$. Let $C = (c(S_1), \dots, c(S_M))'$ be a vector of averaging weights, which is in the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H} = \left\{ C \in [0, 1]^M : \sum_{i=1}^M c(S_i) = 1 \right\}.$$

Ignoring τ_0^2 , which does not depend on the model, the optimal weight vector, C^* , that minimizes the asymptotic MSE is:

$$C^* = \arg \min_{C \in \mathcal{H}} C' A C,$$

where A is an $M \times M$ matrix, whose (i, j) element is given by:

$$A_{[ij]} = w'(I - G_{S_i}) \delta \delta'(I - G_{S_j})' w + w'(G_{S_i} Q G_{S_j}') w.$$

If we replace A with its appropriate estimate, $\hat{A}$, we obtain a feasible estimator:

$$\hat{C} = \arg \min_{C \in \mathcal{H}} C' \hat{A} C. \quad (4)$$

For instance, if we estimate $\delta \delta'$ by $D_n D_n' - \hat{Q}$, then:

$$\hat{A}_{[ij]} = \hat{w}'(I - \hat{G}_{S_i})(D_n D_n' - \hat{Q})(I - \hat{G}_{S_j})' \hat{w} + \hat{w}'(\hat{G}_{S_i} \hat{Q} \hat{G}_{S_j}') \hat{w}.$$

Although there is no closed-form solution for (4), it can be solved numerically by a usual quadratic programming algorithm.

Unfortunately, $\hat{C}$ cannot be a consistent estimator for C^* , because there is no consistent estimator for A . Suppose that $C' \hat{A} C \xrightarrow{d} C' \tilde{A} C$ for a random matrix, $\tilde{A}$, and for all $C \in \mathcal{H}$. Then, we have:

$$\hat{C} \xrightarrow{d} \tilde{C} \equiv \arg \min_{C \in \mathcal{H}} C' \tilde{A} C.$$

Thus, $\hat{C}$ is random, even in the limit.

Let $\hat{c}(S_i)$ and $\tilde{c}(S_i)$ be the i -th element of $\hat{C}$ and $\tilde{C}$, respectively. Furthermore, let $\hat{\mu}_{\text{opt}} = \sum_{i=1}^M \hat{c}(S_i) \hat{\mu}_{S_i}$ denote the averaging estimator using $\hat{C}$. Because $\hat{c}(S_i)$ and $\hat{\mu}_{S_i}$ are both determined through D_n , $\hat{c}(S_i)$ and $\sqrt{n}(\hat{\mu}_{S_i} - \mu_{\text{true}})$ converge jointly to $\tilde{c}(S_i)$ and Λ_{S_i} . Therefore, the limiting distribution of $\hat{\mu}_{\text{opt}}$ is given by:

$$\sqrt{n}(\hat{\mu}_{\text{opt}} - \mu_{\text{true}}) \xrightarrow{d} \sum_{i=1}^M \tilde{c}(S_i) \Lambda_{S_i} \stackrel{d}{=} \Lambda_0 + w' \left(\delta - \sum_{i=1}^M \tilde{c}(S_i) G_{S_i} D \right). \quad (5)$$

Because weights are random, the limiting distribution is no longer normal. Thus, (5) is not readily applicable for inference. However, as suggested by Hjort and Claeskens [21], (5) implies that:

$$\frac{1}{\hat{\kappa}} \left[\sqrt{n}(\hat{\mu}_{\text{opt}} - \mu_{\text{true}}) - \hat{w}' \left(D_n - \sum_{i=1}^M \hat{c}(S_i) \hat{G}_{S_i} D_n \right) \right] \xrightarrow{d} N(0, 1),$$

where $\hat{\kappa}$ is a consistent estimator for $(\tau_0^2 + w' Q w)^{1/2}$. This result can be used to construct a confidence interval for μ_{true} .

5. Example

This section gives a simple example to which our methods are applicable. One of the most popular models described by moment restrictions is the linear instrumental variable model. The full model we consider here is:

$$\begin{aligned} y_i &= x_i' \theta + z_{1i}' \gamma + u_i, \\ E[z_i u_i] &= 0, \end{aligned}$$

where x_i and z_{1i} are $p \times 1$ and $q \times 1$ vectors of explanatory variables. Some elements of x_i are potentially correlated with u_i . The vector of instruments, z_i , is $l \times 1$, which may contain elements of x_i and z_{1i} . Economic theory suggests that x_i should be included in the model, but we are unsure which components of z_{1i} should be included. Thus, the reduced model corresponds to the case that $\gamma = \gamma_0 = (0, \dots, 0)'$.

In this model, J_{full} is given by:

$$J_{\text{full}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} = \begin{pmatrix} E[x_i z_i'] E[z_i z_i' u_i^2]^{-1} E[z_i x_i'] & E[x_i z_i'] E[z_i z_i' u_i^2]^{-1} E[z_i z_{1i}'] \\ E[z_{1i} z_i'] E[z_i z_i' u_i^2]^{-1} E[z_i x_i'] & E[z_{1i} z_i'] E[z_i z_i' u_i^2]^{-1} E[z_i z_{1i}'] \end{pmatrix}.$$

Let $\hat{u}_i$ be the residual from the full model: $\hat{u}_i = y_i - x_i' \hat{\theta}_{\text{full}} - z_{1i}' \hat{\gamma}_{\text{full}}$. Then, for instance, J_{00} can be estimated by:

$$\hat{J}_{00} = \frac{1}{n} \sum_{i=1}^n x_i z_i' \left(\frac{1}{n} \sum_{i=1}^n z_i z_i' \hat{u}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^n z_i x_i'. \quad (6)$$

Other components of J_{full} can be estimated in a similar manner. It also is possible to replace the empirical probability, n^{-1} , with the GEL-induced probability.

If the focus parameter is the k -th element of θ , then we have:

$$\hat{w} = \hat{J}_{10} \hat{J}_{00}^{-1} e_k,$$

where e_k is the k -th unit vector, which have one in the k -th element and zero, elsewhere. On the other hand, if the focus parameter is $\mu(\theta, \gamma) = x'\theta + z_1'\gamma$ for a fixed covariate value $(x', z_1')'$, then:

$$\hat{w} = \hat{J}_{10} \hat{J}_{00}^{-1} x - z_1.$$

To obtain a good estimate of $x'\theta + z_1'\gamma$ for a range of covariate values, rather than a single covariate value, we can utilize the idea of Claeskens and Hjort [36], who address minimizing an averaged risk over the range of covariates, rather than the pointwise risk.

6. Monte Carlo Study

We now investigate the performance of post-selection and averaging estimators by a simple Monte Carlo study. Our EL-based methods are compared with EL-based selection methods of Hong, Preston and Shum [16]. The following post-selection and averaging estimators are considered: (i) AIC-like model selection (ii) BIC-like model selection, (iii) FIC model selection and (iv) an averaging estimator, whose weights are given by (4). AIC- and BIC-like criteria are proposed by Hong, Preston and Shum [16] and are given by:

$$\begin{aligned} \text{AIC}_S &= 2 \sum_{i=1}^n \log(1 - \hat{\tau}'_S m(y_i, \hat{\theta}_S, \hat{\gamma}_S)) - 2(l - p - q_S), \\ \text{BIC}_S &= 2 \sum_{i=1}^n \log(1 - \hat{\tau}'_S m(y_i, \hat{\theta}_S, \hat{\gamma}_S)) - (l - p - q_S) \log n. \end{aligned}$$

We use (6) to estimate J .

We consider the linear instrumental variable model. The DGP is specified by the following equations:

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_i + \theta_2 z_{1i} + \sum_{k=1}^4 \gamma_{kn} z_{k+1,i} + u_i, \\ x_i &= 0.3 z_{6i} + 0.2 z_{7i} + 0.5 u_i, \end{aligned}$$

where $(\theta_0, \theta_1, \theta_2)' = (1, 1, 1)'$ and $(\gamma_{1n}, \gamma_{2n}, \gamma_{3n}, \gamma_{4n})' = \delta / \sqrt{n}$ for some vector $\delta = (\delta_1, \delta_2, \delta_3, \delta_4)'$. Exogenous variables, $z_{1i}, \dots, z_{7i}$, are normally distributed with mean zero and variance one, and the correlation between z_{ki} and z_{li} is $0.5^{|k-l|}$ for $k \neq l$. The vector of instruments is fixed to be $z_i = (1, z_{1i}, \dots, z_{7i})'$. The error term, u_i , is independent of $z_{1i}, \dots, z_{7i}$ and is generated from a standard normal distribution. Thus, the moment restriction for the full model is:

$$E_n \left[z_i \left(y_i - \theta_0 - \theta_1 x_i - \theta_2 z_{1i} - \sum_{k=1}^4 \gamma_{kn} z_{k+1,i} \right) \right] = 0.$$

Table 1. Estimation results; DGP, data generating process; AIC, Akaike information criterion; BIC, Bayesian information criterion; FIC, focused information criterion.

		DGP			
		(1)	(2)	(3)	(4)
Full	Bias	-0.104	-0.109	-0.089	-0.076
	Std	0.544	0.533	0.509	0.489
	RMSE	0.554	0.544	0.516	0.495
Reduced	Bis	-0.279	-0.057	-0.148	-0.048
	Std	0.780	0.473	0.955	0.448
	RMSE	0.828	0.477	0.965	0.450
AIC	Bias	-0.113	-0.099	-0.101	-0.079
	Std	0.559	0.557	0.497	0.509
	RMSE	0.570	0.566	0.507	0.515
BIC	Bias	-0.136	-0.088	-0.104	-0.073
	Std	0.689	0.552	0.499	0.502
	RMSE	0.702	0.559	0.510	0.507
FIC	Bias	-0.139	-0.095	-0.112	-0.076
	Std	0.530	0.509	0.464	0.452
	RMSE	0.548	0.517	0.477	0.458
Averaging	Bias	-0.139	-0.092	-0.107	-0.074
	Std	0.511	0.476	0.455	0.444
	RMSE	0.529	0.484	0.468	0.450

The focus parameter is $\mu = \theta_1$. In many applications, it is often the case that the only parameter of interest in the linear model is the coefficient of the endogenous regressor. Exogenous regressors are included simply to avoid omitted variable bias. Thus, if the bias is small, it may be better to exclude some regressors to reduce the variance. In this simulation, we include the constant term, x_i , and z_{1i} in all candidate models, but some elements of $(z_{2i}, z_{3i}, z_{4i}, z_{5i})'$ may be excluded. That is, some elements of $(\gamma_{1n}, \gamma_{2n}, \gamma_{3n}, \gamma_{4n})'$ are set to zero. Therefore, there are $2^4 = 16$ submodels in total.

To evaluate the performance of the post-selection and averaging estimators, we calculate the bias, standard deviation and root MSE (RMSE) of each estimator over 1,000 repetitions. For reference, we also report the results of the full and reduced models. The sample size is $n = 50$.² We consider four DGPs: (1) $\delta = (1, 1, 1, 1)'$, (2) $\delta = (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})'$, (3) $\delta = (1, \frac{3}{4}, \frac{1}{2}, \frac{1}{4})'$ and (4) $\delta = (\frac{1}{4}, \frac{3}{16}, \frac{1}{8}, \frac{1}{16})'$. The DGPs (1) and (3) are favorable for the full model, while (2) and (4) are favorable for the reduced model. The results are summarized in Table 1.

Table 1 indicates that there are certain cases where we should avoid using the full model, even if it is the correct model. Performance of the full model is poorer than the FIC-selected model for all DGPs.

²Simulations were also conducted for different sample sizes. The results are not reported here, because the difference among candidate models is so small for large n that RMSEs are the almost identical for all models.

As the theory suggests, the efficiency gain of FIC over the full model is large when δ is small. The averaging estimator outperforms all post-selection estimators. It is even better than FIC. As is consistent with findings in the literature, averaging is a useful method to reduce the risk of the estimator.

7. Conclusions

This paper studied GEL-based model selection and averaging methods that are designed to obtain an efficient estimator for the parameter of interest. We modified the local misspecification framework of Claeskens and Hjort [7], so that an FIC can be obtained for moment restriction models. Then, we proposed the averaging estimator by extending the idea of FIC.

In the simulation study, we considered the model selection/averaging problem for the linear instrumental variable model. Although some methods have been advocated for selecting/averaging instruments in the literature, there are few studies on the model selection/averaging problem. The result of the simulation suggests that our averaging can be a useful alternative to existing post-selection estimators.

Acknowledgments

The author thanks Ryo Okui for his comments and suggestions. The author also thanks three referees, seminar participants at the University of Tokyo and participants of a summer workshop on economic theory at Otaru University of Commerce for their comments. The author acknowledges financial support from the Japan Society for the Promotion of Science under KAKENHI 23730215.

References

1. Smith, R.J. Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation. *Econ. J.* **1997**, *107*, 503–519.
2. Newey, W.K.; Smith, R.J. Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* **2004**, *72*, 219–255.
3. Owen, A.B. Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika* **1988**, *75*, 237–249.
4. Qin, J.; Lawless, J. Empirical Likelihood and General Estimating Equations. *Ann. Stat.* **1994**, *22*, 300–325.
5. Kitamura, Y.; Stutzer, M. An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica* **1997**, *65*, 861–874.
6. Imbens, G.W.; Spady, R.H.; Johnson, P. Information Theoretic Approaches to Inference in Moment Condition Models. *Econometrica* **1998**, *66*, 333–357.
7. Claeskens, G.; Hjort, N.L. The Focused Information Criterion. *J. Am. Stat. Assoc.* **2003**, *98*, 900–916.
8. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. Second International Symposium on Information Theory; Petroc, B.; Csake, F., Eds., 1973, pp. 267–281. Akademiai Kiado.
9. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464.

10. Hansen, B.E. Challenges for Econometric Model Selection. *Economet. Theor.* **2005**, *21*, 60–68.
11. Claeskens, G.; Croux, C.; Kerckhoven, J.V. Variable Selection for Logistic Regression Using a Prediction-Focused Information Criterion. *Biometrics* **2006**, *62*, 972–979.
12. Hjort, N.L.; Claeskens, G. Focused Information Criteria and Model Averaging for the Cox Hazard Regression Model. *J. Am. Stat. Assoc.* **2006**, *101*, 1449–1464.
13. Zhang, X.; Liang, H. Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models. *Ann. Stat.* **2011**, *39*, 174–200.
14. Andrews, D.W.; Lu, B. Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models. *J. Econometrics* **2001**, *101*, 123–164.
15. Hansen, L.P. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **1982**, *50*, 1029–1054.
16. Hong, H.; Preston, B.; Shum, M. Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models. *Economet. Theor.* **2003**, *19*, 923–943.
17. Sueishi, N. Information Criteria for Moment Restriction Models. Unpublished Manuscript, Kyoto University, 2013
18. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian Model Averaging: A Tutorial. *Stat. Sci.* **1999**, *14*, 382–417.
19. Buckland, S.T.; Burnham, K.P.; Augustin, N.H. Model Selection: An Integral Part of Inference. *Biometrics* **1997**, *53*, 603–618.
20. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer, 2002.
21. Hjort, N.L.; Claeskens, G. Frequentist Model Average Estimators. *J. Am. Stat. Assoc.* **2003**, *98*, 879–899.
22. Yang, Y. Adaptive Regression by Mixing. *J. Am. Stat. Assoc.* **2001**, *96*, 574–588.
23. Leung, G.; Barron, A.R. Information Theory and Mixing Least-Squares Regressions. *IEEE T. Inform. Theory* **2006**, *52*, 3396–3410.
24. Goldenshluger, A. A Universal Procedure for Aggregating Estimators. *Ann. Stat.* **2009**, *37*, 542–568.
25. Hansen, B.E. Least Squares Model Averaging. *Econometrica* **2007**, *75*, 1175–1189.
26. Wan, A.T.K.; Zhang, X.; Zou, G. Least Squares Model Averaging by Mallows Criterion. *J. Econometrics* **2010**, *156*, 277–283.
27. Hansen, B.E.; Racine, J.S. Jackknife Model Averaging. *J. Econometrics* **2012**, *167*, 38–46.
28. Liu, Q.; Okui, R. Heteroskedasticity-Robust C_p Model Averaging. *Economet. J.* **2013**. Forthcoming.
29. DiTraglia, F.J. Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM. Unpublished Manuscript, University of Pennsylvania, 2012
30. Liu, C.A. A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors. Unpublished Manuscript, National University of Singapore, 2012
31. Martins, L.F.; Gabriel, V.J. Linear Instrumental Variables Model Averaging Estimation. *Comput. Stat. Data An.* **2013**. Forthcoming.

32. Newey, W.K. Generalized Method of Moments Specification Testing. *J. Econometrics* **1985**, *29*, 229–256.
33. Hall, A.R. Hypothesis Testing in Models Estimated by Generalized Method of Moments. In *Generalized Method of Moments Estimation*; Mátyás, L., Ed.; Cambridge University Press, 1999; pp. 75–101.
34. Parente, P.M.; Smith, R.J. GEL Methods for Nonsmooth Moment Indicators. *Economet. Theor.* **2011**, *27*, 74–113.
35. Claeskens, G.; Hjort, N.L. *Model Selection and Model Averaging*; Cambridge University Press, 2008.
36. Claeskens, G.; Hjort, N.L. Minimizing Average Risk in Regression Models. *Economet. Theor.* **2008**, *24*, 493–527.

A. Appendix

This appendix provides a proof for Lemma 3.1. In this appendix, symbols, $\xrightarrow{p}$ and $\xrightarrow{d}$, denote convergence in probability and in distribution with respect to the local sequence, $f_n(y)$.

Let $m_i(\theta, \gamma_S, \gamma_{SC}) = m(y_i, \theta, \gamma_S, \gamma_{SC})$ and $m_i(\theta, \gamma_S) = m(y_i, \theta, \gamma_S, \gamma_{0,SC})$. We define:

$$\tau(\theta, \gamma_S, \gamma_{SC}) = \arg \max_{\tau \in T} E [\rho(\tau' m_i(\theta, \gamma_S, \gamma_{SC}))].$$

Condition 5 implies that $\tau(\theta, \gamma_S, \gamma_{SC})$ is continuous with respect to $(\theta, \gamma_S, \gamma_{SC})$. Moreover:

$$\left. \frac{\partial E[\rho(\tau' m_i(\theta_0, \gamma_{0,S}, \gamma_{0,SC}))]}{\partial \tau} \right|_{\tau=0} = \rho_1(0) E[m_i(\theta_0, \gamma_{0,S}, \gamma_{0,SC})] = 0.$$

Thus, by concavity of $\rho(v)$, $\tau(\theta_0, \gamma_{0,S}, \gamma_{0,SC}) = 0$.

Let $\hat{\tau}_S(\theta, \gamma_S) = \arg \max_{\tau \in T} n^{-1} \sum_{i=1}^n \rho(\tau' m_i(\theta, \gamma_S))$. Then, by construction:

$$\frac{1}{n} \sum_{i=1}^n \rho(\hat{\tau}_S(\theta, \gamma_S)' m_i(\theta, \gamma_S)) \geq \frac{1}{n} \sum_{i=1}^n \rho(\tau(\theta, \gamma_S, \gamma_{0,SC})' m_i(\theta, \gamma_S)).$$

Also, let $L = E[\rho(\tau(\theta_0, \gamma_{0,S}, \gamma_{0,SC})' m_i(\theta_0, \gamma_{0,S}))] = \rho(0)$. Then, Condition 6 and the saddle-point property imply that:

$$E[\rho(\tau(\theta, \gamma_S, \gamma_{0,SC})' m_i(\theta, \gamma_S))] > L$$

for $\theta \neq \theta_0$ and $\gamma_S \neq \gamma_{0,S}$. Let $B(\theta_0, \gamma_{0,S}, \epsilon)$ be an open ball of radius, ϵ , around $(\theta_0, \gamma_{0,S})$. Conditions 1–4 imply:

$$\left| \frac{1}{n} \sum_{i=1}^n \rho(\tau(\theta, \gamma_S, \gamma_{0,SC})' m_i(\theta, \gamma_S)) - E[\rho(\tau(\theta, \gamma_S, \gamma_{0,SC})' m_i(\theta, \gamma_S))] \right| \xrightarrow{p} 0$$

uniformly over $\theta \in \Theta$ and $\gamma_S \in \Gamma_S$. Thus, for any $\epsilon > 0$, there exists $\delta > 0$, such that:

$$P_n \left(\inf_{(\theta, \gamma_S) \in \Theta \times \Gamma_S \setminus B(\theta_0, \gamma_{0,S}, \epsilon)} \frac{1}{n} \sum_{i=1}^n \rho(\hat{\tau}_S(\theta, \gamma_S)' m_i(\theta, \gamma_S)) < L + \delta \right) \rightarrow 0, \quad (7)$$

where P_n is the probability under $f_n(y)$. Conditions 1–4 also imply $\hat{\tau}_S(\theta_0, \gamma_{0,S}) \xrightarrow{p} \tau(\theta_0, \gamma_{0,S}) = 0$. Therefore, we obtain:

$$P_n \left(\frac{1}{n} \sum_{i=1}^n \rho(\hat{\tau}_S(\theta_0, \gamma_{0,S})' m_i(\theta_0, \gamma_{0,S})) > L + \delta \right) \rightarrow 0. \quad (8)$$

Combining (7) and (8), we have $\hat{\theta}_S \xrightarrow{p} \theta_0$ and $\hat{\gamma}_S \xrightarrow{p} \gamma_{0,S}$. Moreover, we have:

$$\left| \frac{1}{n} \sum_{i=1}^n \rho(\tau' m_i(\hat{\theta}_S, \hat{\gamma}_S)) - E[\rho(\tau' m_i(\theta_0, \gamma_{0,S}))] \right| \xrightarrow{p} 0$$

uniformly over $\tau \in \mathcal{T}$. Thus, $\hat{\tau}_S = \hat{\tau}_S(\hat{\theta}_S, \hat{\gamma}_S) \xrightarrow{p} \tau(\theta_0, \gamma_{0,S}, \gamma_{0,S^c}) = 0$.

Next, we derive the asymptotic distribution. The first-order conditions for $(\hat{\theta}'_S, \hat{\gamma}'_S, \hat{\tau}'_S)'$ are:

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^n \rho_1(\hat{\tau}'_S m_i(\hat{\theta}_S, \hat{\gamma}_S)) m_{\theta i}(\hat{\theta}_S, \hat{\gamma}_S)' \hat{\tau}_S, \\ 0 &= \frac{1}{n} \sum_{i=1}^n \rho_1(\hat{\tau}'_S m_i(\hat{\theta}_S, \hat{\gamma}_S)) m_{\gamma_S i}(\hat{\theta}_S, \hat{\gamma}_S)' \hat{\tau}_S, \\ 0 &= \frac{1}{n} \sum_{i=1}^n \rho_1(\hat{\tau}'_S m_i(\hat{\theta}_S, \hat{\gamma}_S)) m_i(\hat{\theta}_S, \hat{\gamma}_S), \end{aligned}$$

where $m_{\theta i}(\theta, \gamma_S) = \frac{\partial}{\partial \theta} m_i(\theta, \gamma_S)$ and $m_{\gamma_S i}(\theta, \gamma_S) = \frac{\partial}{\partial \gamma_S} m_i(\theta, \gamma_S)$. By Condition 11 and consistency of the estimator, expanding the first-order conditions around $(\theta_0, \gamma_{0,S}, 0)$, we obtain:

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_S - \theta_0 \\ \hat{\gamma}_S - \gamma_{0,S} \\ \hat{\tau}_S \end{pmatrix} = - \begin{pmatrix} 0 & 0 & E_n[m_{\theta i}]' \\ 0 & 0 & E_n[m_{\gamma_S i}]' \\ E_n[m_{\theta i}] & E_n[m_{\gamma_S i}] & E_n[m_i m_i'] \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n m_i \end{pmatrix} + o_p(1).$$

Let $w_{in} = \eta'(m_i - E_n[m_i])$, where η is any $l \times 1$ vector, such that $\eta'\eta = 1$. Then, we have:

$$\sigma_n^2 \equiv E_n[w_{in}^2] = \eta' E_n[m_i m_i'] \eta + o(1/n)$$

and:

$$\begin{aligned} \frac{1}{\sigma_n^2} E_n[w_{in}^2 1\{|w_i| \geq \sqrt{n} \sigma_n \epsilon\}] &\leq E_n \left[\frac{|w_{in}|^{2+\alpha}}{|w_{in}|^\alpha} 1\{|w_{in}| \geq \sqrt{n} \sigma_n \epsilon\} \right] \\ &\leq \frac{1}{n^{\alpha/2} \sigma_n^{2+\alpha} \epsilon^\alpha} E_n[|w_{in}|^{2+\alpha}] \rightarrow 0 \end{aligned}$$

by Condition 9. Thus, by the Lindeberg-Feller Theorem and Condition 13:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n w_{in} \xrightarrow{d} N(0, \eta' E[m_i m_i'] \eta).$$

Furthermore, by Condition 12, we have:

$$\sqrt{n} E_n[m_i] = -E[m_{\gamma_i}] \delta + o(1).$$

Therefore, by the Cramer-Wold device, we obtain:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n m_i \xrightarrow{d} N(-E[m_{\gamma_i}]\delta, E[m_i m_i']),$$

which implies the desired result.

© 2013 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).