Statistica Sinica

A UNIFIED APPROACH TO FOCUSED INFORMATION CRITERION AND PLUG-IN AVERAGING METHOD

Xinyu Zhang[1,2] and Chu-An Liu[3,*]

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

[2]*Beijing Academy of Artificial Intelligence*

[3]*Institute of Economics, Academia Sinica*

Abstract: Unlike the traditional model selection criterion, which picks a single model based on the global fit of the model, the focused information criterion proposed by Claeskens and Hjort (2003) is tailored to the parameter of interest and aims to select a model based on the parameter under focus. In this paper, we develop a focused information criterion and a plug-in averaging method for a general class of estimators in a unified theoretical framework, and investigate their asymptotic and finite sample properties. Monte Carlo simulations and real data analysis show that both proposed selection and averaging methods compare favorably with other methods.

Key words and phrases: Focused information criterion, Model averaging, Model selection.

*Corresponding author: Chu-An Liu, Institute of Economics, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei, 115, Taiwan. Email: caliu@econ.sinica.edu.tw.

1. Introduction

There is a long history of model selection methods in the econometric and statistical literature. The traditional model selection criteria such as the Akaike information criterion and Bayesian information criterion aim to choose one single model based on its global fit. The selected model provides the best approximation to the unknown true data generating process, but it may not be ideal for estimating a specific model parameter under consideration. For example, Hansen (2005) showed that finite-sample optimal model selection might be quite sensitive to the choice of parameter of interest. Claeskens, Croux, and Van Kerckhoven (2006) gave some specific examples in biostatistics in which no single model is good for every patient subgroup. Instead of choosing one single model to explain all aspects of data, the focused information criterion (FIC; Claeskens and Hjort, 2003) aims to select a model based on the parameter under focus, and allows different models to be chosen for different parameters of interest.

Since the seminal work of Claeskens and Hjort (2003), the FIC has been investigated in different models, including the Cox hazard regression model (Hjort and Claeskens, 2006), the general semiparametric model (Claeskens and Carroll, 2007), the generalized additive partial linear model (Zhang and Liang, 2011), the varying-coefficient partially linear measurement er-

ror model (Wang, Zou, and Wan, 2012), the Tobin model with a nonzero threshold (Zhang, Wan, and Zhou, 2012), the partially linear single-index model (Yu et al., 2013), the linear mixed-effects model (Chen, Zou, and Zhang, 2013), generalized empirical likelihood estimation (Sueishi, 2013), the graphical model (Pircalabelu, Claeskens, and Waldorp, 2015), propensity score weighted estimation of the treatment effects (Lu, 2015; Kitagawa and Muris, 2016), the choice between parametric and nonparametric models (Jullum and Hjort, 2017), generalized method of moments estimation (DiTraglia, 2016; Chang and DiTraglia, 2018), vector autoregressive models (Lohmeyer et al., 2019), and others. It is well known that many of these estimators share a common structure, which is useful in deriving the FIC in different model setups. Therefore, it would be interesting to know whether it is feasible to develop the FIC for various models in a unified theoretical framework instead of in a case-by-case manner.

In this paper, we develop the FIC for a general class of estimators, referred to as extremum estimators by Newey and McFadden (1994), that maximizes the sample objective function. The goal is to evaluate and select a model based on the parameter under focus in a general setting. We first extend the asymptotic theory of extremum estimators for drifting sequences of parameters, and demonstrate that the trade-off between bias and variance

remains in the asymptotic theory. We then follow Claeskens and Hjort (2003) and propose the FIC for extremum estimators. The proposed FIC is an asymptotically unbiased estimator of the asymptotic mean squared error (AMSE) for the limiting distribution of the focus parameter estimate. Thus, the FIC aims to choose the model that achieves the minimum estimated AMSE. We apply our results to several examples and provide the FIC in each case, including the nonlinear least squares estimator, the maximum likelihood estimator, the generalized method of moments estimator, and the minimum distance estimator.

As an alternative to model selection, a model averaging estimator incorporates all available information and constructs a weighted average of the estimates across all potential models. There are two main model averaging methods: Bayesian model averaging and frequentist model averaging; see Hoeting et al. (1999), Claeskens and Hjort (2008), Moral-Benito (2015), and Steel (2020) for a literature review. In this paper, we propose a plug-in averaging method with data-driven weights for extremum estimators. We first derive the limiting distribution of the averaging estimator with fixed weights for the parameter under focus, and use this asymptotic result to characterize the optimal weights of the averaging estimator under the quadratic loss function. We then propose a plug-in method to estimate the

infeasible optimal weights, and use these estimated weights to construct a frequentist model averaging estimator of the focus parameter.

We investigate the asymptotic and finite sample properties of the proposed FIC and plug-in averaging method. We show that both the FIC and estimated weights are asymptotically random under the local asymptotic framework, and hence the FIC model selection estimator and the averaging estimator with data-driven weights have nonstandard asymptotic distributions. We use a simple three-nested-model framework to illustrate the effect of the estimated local parameter on asymptotic behavior of the FIC and plug-in averaging method. In simulations, we compare the finite sample performance of the FIC and plug-in averaging method with other existing model selection and model averaging methods. In real data analysis, we apply the proposed methods to investigate the relationship between income and education. Both simulation studies and empirical results show that the proposed methods perform well and generally achieve lower mean squared errors than other methods.

The rest of the paper is organized as follows. Section 2 presents the model, extremum estimators, and the asymptotic framework. Section 3 introduces the FIC and plug-in averaging method for the extremum estimator and studies their asymptotic behavior. Section 4 evaluates the asymptotic

and finite sample performance of the proposed methods. Section 5 concludes the paper. Proofs are included in the supplementary materials.

2. Model framework and estimation

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')' \in \boldsymbol{\Theta} \subset \mathbb{R}^{p+q}$ denote a $p+q$ vector of unknown parameters, where $\boldsymbol{\Theta}$ is the set of possible parameter values. Suppose we have a sample objective function $\widehat{Q}_n(\boldsymbol{\theta})$ that depends on data and sample size n, and we consider a general class of estimators, referred to as extremum estimators by Newey and McFadden (1994), that maximizes this objective function. Notice that $\widehat{Q}_n(\boldsymbol{\theta})$ could be a negative log-likelihood function, a least squares function, a minimum-distance criterion function, and so on. For example, if we set $\widehat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\mathbf{x}_i)$, where $m_{\boldsymbol{\theta}}(\cdot)$ is a real-value function of $\mathbf{x}_i$, then the extremum estimator is an M-estimator, which includes the maximum likelihood estimator and nonlinear least squares estimator as special cases. If we set $\widehat{Q}_n(\boldsymbol{\theta}) = -g_n(\boldsymbol{\theta})'\mathbf{W}_n g_n(\boldsymbol{\theta})$, where $\mathbf{W}_n$ is a positive semi-definite weight matrix and $g_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i, \boldsymbol{\theta})$ is a sample average of moment functions, then the extremum estimator is the generalized method of moments estimator.

Our goal is to select a model based on the parameter under focus in a general setting that allows for parameter uncertainty. In our framework,

the candidate models could be nested or non-nested, and in each candidate model, we are uncertain about which model parameters should be included in the model. Without loss of generality, we assume that $\boldsymbol{\beta}$ is a $p \times 1$ vector of "must-have" parameters that must be included in the model based on theoretical grounds, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of "potentially relevant" parameters that may or may not be included in the model. Consider a sequence of submodels indexed by $s = 1, \ldots, S$, where the sth submodel includes all $\boldsymbol{\beta}$ but some or none of the components $\boldsymbol{\gamma}$. Since the true value of $\boldsymbol{\gamma}$ could be zeros, we could restrict some elements of $\boldsymbol{\gamma}$ zeros to obtain candidate models and allow for the parameter uncertainty. If we consider a sequence of nested models, then we have $S = q + 1$ submodels. If we consider all possible subsets of potentially relevant parameters $\boldsymbol{\gamma}$, then we have $S = 2^q$ submodels.

For the full model, we include all $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, while for the narrow model, we only include $\boldsymbol{\beta}$ and set all $\boldsymbol{\gamma}$ to be zeros. We could also set some $\boldsymbol{\gamma}$ zeros and consider an intermediate model between the full model and the narrow model. Let $\boldsymbol{\gamma}_s$ denote the included elements of $\boldsymbol{\gamma}$ in the sth submodel, and $\boldsymbol{\gamma}_{s^c}$ the remaining elements of $\boldsymbol{\gamma}$ in the sth submodel. For the full model, the unknown parameters are $\boldsymbol{\theta}$, and the extremum estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widehat{Q}_n(\boldsymbol{\theta}). \tag{2.1}$$

For the sth submodel, the unknown parameters are $\boldsymbol{\eta}_s = (\boldsymbol{\beta}', \boldsymbol{\gamma}_s')'$. Let $\boldsymbol{\Pi}_s$ be a $(p+q_s)\times(p+q)$ projection matrix such that $\boldsymbol{\Pi}_s\boldsymbol{\theta} = \boldsymbol{\eta}_s$, where q_s is the dimension of $\boldsymbol{\gamma}_s$. Similarly, let $\boldsymbol{\Pi}_{s^c}$ be a projection matrix such that $\boldsymbol{\Pi}_{s^c}\boldsymbol{\theta} = \boldsymbol{\gamma}_{s^c}$. Hence, we can write $\widehat{Q}_n(\boldsymbol{\theta})$ as $\widehat{Q}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_s, \boldsymbol{\gamma}_{s^c})$ and the extremum estimator for the sth submodel is

$$\widehat{\boldsymbol{\theta}}_s = \boldsymbol{\Pi}_s'\widehat{\boldsymbol{\eta}}_s = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widehat{Q}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_s, \mathbf{0}), \tag{2.2}$$

where $\mathbf{0}$ is a zero vector. Note that $\widehat{\boldsymbol{\theta}}_s$ is a $(p+q)\times 1$ vector with the values of $\boldsymbol{\gamma}_{s^c}$ being zero.

We now state the regularity conditions required for asymptotic results, where all limiting processes here and throughout the text are with respect to $n\to\infty$. Suppose that the objective function $\widehat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$, and $Q_0(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$. Define $\boldsymbol{\theta}_0^* = (\boldsymbol{\beta}_0', \mathbf{0}')'$ as the null points. Let

$$\mathbf{H}_n(\boldsymbol{\theta}) = \frac{\partial^2 \widehat{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \text{ and } \mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 Q_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

be the Hessian matrix of second derivatives and the expected Hessian matrix, respectively. Let $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and convergence in distribution, respectively. Let $\|\cdot\|$ denote the Euclidean norm.

Assumption 1. *(i)* $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \xrightarrow{p} \mathbf{0}$. *(ii)* $\boldsymbol{\theta}_0$ *is in the interior of* $\boldsymbol{\Theta}$. *(iii)* $\widehat{Q}_n(\boldsymbol{\theta})$ *is twice continuously differentiable in a neighborhood* $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ *of* $\boldsymbol{\theta}_0$. *(iv)* $\sqrt{n}\frac{\partial}{\partial \boldsymbol{\theta}}\widehat{Q}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$. *(v) There is* $\mathbf{H}(\boldsymbol{\theta})$ *that is continuous at* $\boldsymbol{\theta}_0$ *for every* n, *and* $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{H}_n(\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})\| \xrightarrow{p} \mathbf{0}$. *(vi)* $\mathbf{H}(\boldsymbol{\theta}_0)$ *is nonsingular and negative definite.*

Assumption 1 is identical to conditions in Theorem 3.1 of Newey and McFadden (1994). Assumption 1(i) assumes the consistency of $\widehat{\boldsymbol{\theta}}$, and this condition holds under appropriate primitive assumptions; see the discussions in Section 2 of Newey and McFadden (1994). Let $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}_0)$. Under Assumption 1, Theorem 3.1 of Newey and McFadden (1994) demonstrates the asymptotic normality of $\widehat{\boldsymbol{\theta}}$:

$$\mathbf{Z}_n \equiv \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{Z} \sim N(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}), \tag{2.3}$$

where $\mathbf{Z}$ is a normal random vector and $\boldsymbol{\Sigma}$ is a positive definite matrix.

Assumption 2. $\widehat{Q}_n(\boldsymbol{\theta})$ *is three times differentiable in a neighborhood* $\boldsymbol{\Theta}_0^* \subset \boldsymbol{\Theta}$ *of* $\boldsymbol{\theta}_0^*$, *and the third partial derivative of* $\widehat{Q}_n(\boldsymbol{\theta})$ *satisfies*

$$\sup_{\boldsymbol{\theta}_0^* \in \boldsymbol{\Theta}_0^*} \frac{\partial^3 \widehat{Q}_n(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} = o_p(n^{1/2}).$$

Assumption 2 requires that the third partial derivative of the objective function is bounded by $n^{1/2}$. This condition holds for most models, and it

is similar to Condition C4 in Hjort and Claeskens (2003) and Condition A4 in Claeskens and Carroll (2007). The quantile regression model, however, is excluded from our framework due to the failure of differentiability. For the focused information criterion in the quantile regression framework, see Behl, Claeskens, and Dette (2014) and Xu, Wang, and Huang (2014).

Assumption 3. $\boldsymbol{\gamma}_0 \equiv \boldsymbol{\gamma}_{0,n} = \boldsymbol{\delta}_0/\sqrt{n}$ *where* $\boldsymbol{\delta}_0$ *is an unknown constant vector.*

Assumption 3 specifies that $\boldsymbol{\gamma}_0$ is in a local $n^{-1/2}$ neighborhood of zero, and thus $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\delta}_0'/\sqrt{n})'$. This is a technique to ensure that the asymptotic mean squared error of each submodel estimator remains finite. The local asymptotic framework is a technical device commonly used to analyze the asymptotic and finite sample properties of the model selection estimator, for example as in Hjort and Claeskens (2003), Leeb and Pötscher (2005), and Claeskens and Hjort (2008). This assumption implies that all of the submodels are close to each other as the sample size increases. The assumption also has an advantage of yielding the same stochastic order of squared biases and variances. Hence, the optimal model is the one that achieves the best trade-off between bias and variance in this context. Alternatively, other works use the assumption that the parameters decay in an appropriate rate such that the squared biases and variances have the same

order; for example, see Hansen (2007) and Cheng, Ing, and Yu (2015).

In the standard asymptotics with fixed parameters setup, the model bias tends to infinity with the sample size, and hence the asymptotic approximations break down. To obtain a useful approximation, we study perturbations of the model with the parameters $\boldsymbol{\gamma}$ being a local neighborhood of zero. Let $\mathbf{I}$ denote an identity matrix. The following theorem presents the asymptotic distribution of the extremum estimator for each submodel in the local asymptotic framework.

Theorem 1. *Suppose that Assumptions 1–3 hold. As $n \to \infty$, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) \xrightarrow{d} \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z} \sim N(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0, \mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_s}), \quad (2.4)$$

where $\mathbf{H}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{H}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$ *and* $\boldsymbol{\Pi}_0 = (\mathbf{0}_{q\times p}, \mathbf{I}_q)'$.

Remark 1. Theorem 1 extends the asymptotic theory of extremum estimators for drifting sequences of parameters, and it implies that the submodel estimator $\widehat{\boldsymbol{\theta}}_s$ is root-n consistent. When we set $\boldsymbol{\Pi}_s = \mathbf{I}_{p+q}$ for the full model, we have $\widehat{\boldsymbol{\theta}}_s = \widehat{\boldsymbol{\theta}}$. In this case, our result (2.4) is simplified to the asymptotic distribution of the full model estimator presented in (2.3), which corresponds to Theorem 3.1 of Newey and McFadden (1994). Here, $\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0$ and $\mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_s}$ represent the asymptotic bias and the asymptotic variance of the submodel estimator. Our theorem demonstrates that the trade-off

between squared biases and variances remains in the asymptotic theory, and this feature is essential for the FIC and plug-in averaging method.

Remark 2. The proof of Theorem 1 is not a trivial extension of the already existing results. Notice that we impose the condition that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ instead of the condition that $\widehat{\boldsymbol{\theta}}_s \xrightarrow{p} \boldsymbol{\theta}_0^*$. The former condition is imposed on the full model only, but the latter condition is imposed on all candidate models. To derive the asymptotic distribution of the submodel estimator $\widehat{\boldsymbol{\theta}}_s$, we first adopt a similar strategy in Fan and Li (2001) and Wang and Leng (2007) and show that $\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2})$. We then show that $\widehat{\boldsymbol{\theta}}_s$ is approximatively a linear function of $\widehat{\boldsymbol{\theta}}$ as follows

$$\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^* = \widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^*) + o_p(n^{-1/2}), \tag{2.5}$$

where $\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\widehat{\mathbf{H}}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$ and $\widehat{\mathbf{H}} = \mathbf{H}_n(\widehat{\boldsymbol{\theta}})$. Thus, if we multiply both sides of (2.5) by $\sqrt{n}$, the first term converges to a normal distribution by (2.3) and Slutsky's theorem, and the second term converges to an asymptotic bias by Assumption 3. Thus, we demonstrate that the asymptotic distribution of the submodel estimator is a linear function of the normal random vector $\mathbf{Z}$.

3. Focused information criterion and plug-in averaging method

In this section, we propose a focused information criterion for extremum estimators. As an illustration, we apply the general results to the nonlinear least squares (NLS) estimator. We also provide additional examples to illustrate the general results in the supplementary materials, including the maximum likelihood estimator, the generalized method of moments estimator, and the minimum distance estimator. We next extend the idea of the FIC from model selection to model averaging and develop a plug-in averaging method for extremum estimators. In the last subsection, we study the asymptotic behavior of the FIC and plug-in averaging method.

3.1 The FIC for extremum estimators

Empirical studies tend to focus on one particular parameter instead of assessing the overall properties of the model. Unlike the traditional model selection approaches, which assess the global fit of the model, we evaluate the model based on the parameter under focus. Let $\mu = \mu(\boldsymbol{\theta}) = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$ be a focus parameter, which is a smooth real-valued function. Notice that if μ depends only on $\boldsymbol{\gamma}$ and the estimator in a model that set $\boldsymbol{\gamma} = \mathbf{0}$, then Assumption 1 (ii) does not hold. This is because the set $\boldsymbol{\Theta}$ includes only one point $\boldsymbol{\gamma} = \mathbf{0}$, and there is no interior in the set $\boldsymbol{\Theta}$. Let $\mu_0 = \mu(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\beta}_0, \boldsymbol{\delta}_0/\sqrt{n})$

be the focus parameter evaluated at $\boldsymbol{\theta}_0$. For the sth submodel, μ_0 is estimated by $\widehat{\mu}_s = \mu(\widehat{\boldsymbol{\theta}}_s)$. Assume that the partial derivatives of $\mu(\boldsymbol{\theta})$ are continuous in a neighborhood of $\boldsymbol{\theta}_0^*$. Let $\mathbf{D}_{\boldsymbol{\theta}} = (\mathbf{D}_{\boldsymbol{\beta}}', \mathbf{D}_{\boldsymbol{\gamma}}')'$ be partial derivatives evaluated at the null points $\boldsymbol{\theta}_0^*$, that is,

$$\mathbf{D}_{\boldsymbol{\beta}} = \frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \quad \text{and} \quad \mathbf{D}_{\boldsymbol{\gamma}} = \frac{\partial \mu(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*}.$$

We aim to select a model with the lowest possible AMSE of $\widehat{\mu}_s$ under the quadratic loss function. We first derive the asymptotic distribution of $\widehat{\mu}_s$ for each submodel in the local asymptotic framework, and then define the AMSE of $\widehat{\mu}_s$ as the squared bias plus the variance of the asymptotic distribution.

Corollary 1. *Suppose that Assumptions 1–3 hold. As $n \to \infty$, we have*

$$\begin{aligned}\sqrt{n}(\widehat{\mu}_s - \mu_0) \xrightarrow{d} \Lambda_s &\equiv \mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z} \\ &\sim N(\mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0, \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{D}_{\boldsymbol{\theta}}). \quad (3.1)\end{aligned}$$

From Corollary 1, a direct calculation yields

$$\begin{aligned}\mathrm{E}(\Lambda_s^2) = &\mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0\boldsymbol{\delta}_0'\boldsymbol{\Pi}_0'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})'\mathbf{D}_{\boldsymbol{\theta}} \\ &+ \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{D}_{\boldsymbol{\theta}}. \quad (3.2)\end{aligned}$$

Since $\mathbf{D}_{\boldsymbol{\theta}}$ depends on the focus parameter μ, we can use (3.2) to select a proper submodel depending on the parameter of interest. To use (3.2) for

model selection, we need to replace the unknown parameters $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\delta}_0$ with the sample analogs. The proposed FIC of the sth submodel is defined as

$$\begin{aligned}\mathrm{FIC}_s = &\widehat{\mathbf{D}}_{\boldsymbol{\theta}}'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\boldsymbol{\Pi}_0'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}} - \mathbf{I}_{p+q})'\widehat{\mathbf{D}}_{\boldsymbol{\theta}} \\ &+ \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \end{aligned} \tag{3.3}$$

which is an asymptotically unbiased estimator of the mean squared error $\mathrm{E}(\Lambda_s^2)$ in the sense that the mean of the asymptotic distribution of FIC_s equals the mean squared error $\mathrm{E}(\Lambda_s^2)$. Here, $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is defined in the following (3.5). In practice, we select the model with the lowest value of FIC_s.

We now discuss the sample analog estimators in (3.3). We first consider the estimators in the second term of (3.3). Recall that $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\gamma}}')'$ is the extremum estimator from the full model. Define $\widehat{\mathbf{D}}_{\boldsymbol{\theta}} = \partial\mu(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*}$, where $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\beta}}', \mathbf{0}')'$. Since $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ by (2.3), it follows that $\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$ is a consistent estimator of $\mathbf{D}_{\boldsymbol{\theta}}$. For the covariance matrix $\mathbf{H}$, we can consistently estimate $\mathbf{H}$ by the sample analog $\widehat{\mathbf{H}}$ under Assumption 1. Similarly, the covariance matrix $\boldsymbol{\Sigma}$ can also be consistently estimated by the sample analog $\widehat{\boldsymbol{\Sigma}}$.

We now consider the estimator for the local parameter $\boldsymbol{\delta}_0$. Unlike $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$, the consistent estimator for $\boldsymbol{\delta}_0$ is not available due to the local asymptotic framework. We can, however, construct an asymptotically un-

biased estimator of $\boldsymbol{\delta}_0$ by using the extremum estimator from the full model. The asymptotically unbiased estimator of $\boldsymbol{\delta}_0$ is defined as $\widehat{\boldsymbol{\delta}} = \sqrt{n}\widehat{\boldsymbol{\gamma}}$. From (2.3) and Assumption 3, we can show that

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 = \sqrt{n}\boldsymbol{\Pi}_0'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Pi}_0'\mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}\boldsymbol{\Pi}_0). \tag{3.4}$$

As shown above, $\widehat{\boldsymbol{\delta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\delta}_0$. Therefore, the asymptotically unbiased estimator of $\boldsymbol{\delta}_0\boldsymbol{\delta}_0'$ is

$$\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} = \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \boldsymbol{\Pi}_0'\widehat{\mathbf{H}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{H}}^{-1}\boldsymbol{\Pi}_0. \tag{3.5}$$

3.2 Example: Nonlinear least squares estimator

Suppose the data $(y_i, \mathbf{x}_i')'$ are i.i.d. Consider a nonlinear regression model

$$y_i = h(\mathbf{x}_i, \boldsymbol{\theta}_0) + e_i, \tag{3.6}$$

where $\boldsymbol{\theta}_0$ is a vector of unknown parameters, the parametric regression function $h(\mathbf{x}_i, \boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$, and e_i is an unobservable regression error with $\mathrm{E}(e_i|\mathbf{x}_i) = 0$. If $h(\mathbf{x}_i, \boldsymbol{\theta}_0) = \mathbf{x}_i'\boldsymbol{\theta}_0$, then we have the classical linear regression model. The NLS estimator $\widehat{\boldsymbol{\theta}}$ maximizes the following objective function

$$\widehat{Q}_n(\boldsymbol{\theta}) = -\frac{1}{2n}\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i, \boldsymbol{\theta}))^2, \tag{3.7}$$

where $1/2$ is a scale factor that has no effect on the asymptotic results. Note that maximizing $\widehat{Q}_n(\boldsymbol{\theta})$ is equivalent to minimizing the sum of squared

errors $S_n(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - h(\mathbf{x}_i, \boldsymbol{\theta}))^2$. Here the objective function $\widehat{Q}_n(\boldsymbol{\theta})$ converges to $Q_0(\boldsymbol{\theta}) = \mathrm{E}(y_i - h(\mathbf{x}_i, \boldsymbol{\theta}))^2/2$. Thus,

$$\mathbf{H}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}_i, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}) - (y_i - h(\mathbf{x}_i, \boldsymbol{\theta})) \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}) \right), \tag{3.8}$$

$$\mathbf{H}(\boldsymbol{\theta}) = -\mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}_i, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}) \right) + \mathrm{E}\left((y_i - h(\mathbf{x}_i, \boldsymbol{\theta})) \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}) \right), \tag{3.9}$$

and

$$\boldsymbol{\Sigma} = \mathrm{E}\left(e_i^2 \frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}_i, \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}_0) \right). \tag{3.10}$$

From (3.6) and (3.9), we have $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}_0) = -\mathrm{E}(\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}_i, \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta}_0))$. By Theorem 1, it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) \xrightarrow{d} \mathbf{H}_{\boldsymbol{\Pi}_s} \mathbf{H} (\mathbf{Z} + \boldsymbol{\Pi}_0 \boldsymbol{\delta}_0) \sim N(\mathbf{H}_{\boldsymbol{\Pi}_s} \mathbf{H} \boldsymbol{\Pi}_0 \boldsymbol{\delta}_0, \mathbf{V}_{\boldsymbol{\Pi}_s}), \tag{3.11}$$

where $\mathbf{V}_{\boldsymbol{\Pi}_s} = \mathbf{H}_{\boldsymbol{\Pi}_s} \boldsymbol{\Sigma} \mathbf{H}_{\boldsymbol{\Pi}_s}$ and $\mathbf{H}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s' (\boldsymbol{\Pi}_s \mathbf{H} \boldsymbol{\Pi}_s')^{-1} \boldsymbol{\Pi}_s$. In the supplementary materials, we verify the high-level assumptions for the NLS estimator. Thus, by Corollary 1, the FIC for the NLS estimator is defined as

$$\begin{aligned} \mathrm{FIC}_s = {} & \widehat{\mathbf{D}}_{\boldsymbol{\theta}}' (\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s} \widehat{\mathbf{H}} - \mathbf{I}_{p+q}) \boldsymbol{\Pi}_0 \widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} \boldsymbol{\Pi}_0' (\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s} \widehat{\mathbf{H}} - \mathbf{I}_{p+q})' \widehat{\mathbf{D}}_{\boldsymbol{\theta}} \\ & + \widehat{\mathbf{D}}_{\boldsymbol{\theta}}' \widehat{\mathbf{V}}_{\boldsymbol{\Pi}_s} \widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \end{aligned} \tag{3.12}$$

where $\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$, $\widehat{\mathbf{H}}$, and $\widehat{\boldsymbol{\Sigma}}$ are the sample analogs of $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$, and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is the asymptotically unbiased estimator of $\boldsymbol{\delta}_0 \boldsymbol{\delta}_0'$.

When the error term e_i is homoskedastic, i.e., $\mathrm{E}(e_i^2|\mathbf{x}_i) = \sigma^2$, we have $\boldsymbol{\Sigma} = -\sigma^2\mathbf{H}$, and the covariance matrix $\mathbf{V}_{\boldsymbol{\Pi}_s}$ is simplified as $-\sigma^2\mathbf{H}_{\boldsymbol{\Pi}_s}$. In this case, the FIC for the NLS estimator is defined as

$$\begin{aligned}\mathrm{FIC}_s = {} & \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\boldsymbol{\Pi}_0'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{H}} - \mathbf{I}_{p+q})'\widehat{\mathbf{D}}_{\boldsymbol{\theta}} \\ & - \widehat{\sigma}^2\widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \end{aligned} \tag{3.13}$$

where $\widehat{\sigma}^2$ is the sample analog of σ^2.

3.3 Plug-in averaging method

In this section, we extend the idea of the FIC to the averaging estimator and develop a plug-in averaging method for extremum estimators. We first introduce the averaging estimator of the focus parameter. Let $w_s \geq 0$ be the weight corresponding to the sth submodel and $\mathbf{w} = (w_1, \ldots, w_S)'$ be a weight vector belonging to the weight set $\mathcal{W} = \{\mathbf{w} \in [0,1]^S : \sum_{s=1}^S w_s = 1\}$. That is, the weight vector lies in the unit simplex in $\mathbb{R}^S$. The model averaging estimator of μ_0 is defined as

$$\widehat{\mu}(\mathbf{w}) = \sum_{s=1}^{S} w_s\widehat{\mu}_s. \tag{3.14}$$

Note that the model selection estimator based on the information criterion is a special case of the model averaging estimator. The FIC proposed in (3.3) puts the whole weight on the model with the smallest value of the

FIC_s and gives other models zero weights. Thus, the weight function of the FIC is $\widehat{w}_s = \mathbf{1}\{\text{FIC}_s = \min(\text{FIC}_1, \text{FIC}_2, \ldots, \text{FIC}_S)\}$, where $\mathbf{1}\{\cdot\}$ is an indicator function that takes a value of either 0 or 1.

We now consider a general weight function instead of a zero-one weight function. Instead of comparing the AMSE of each submodel, we first derive the AMSE of the averaging estimator with fixed weight in a local asymptotic framework. Next, we use this asymptotic result to characterize the optimal weights of the averaging estimator under the quadratic loss function. We then follow Wan, Zhang, and Wang (2014) and Liu (2015) and propose a plug-in method to estimate the infeasible optimal weights. The following theorem presents the asymptotic distribution of the averaging estimator with fixed weights.

Theorem 2. *Suppose that Assumptions 1–3 hold. As $n \to \infty$, we have*

$$\sqrt{n}(\widehat{\mu}(\mathbf{w}) - \mu_0) \xrightarrow{d} N(\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{B}(\mathbf{w})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0, V(\mathbf{w})), \tag{3.15}$$

where

$$\mathbf{B}(\mathbf{w}) = \sum_{s=1}^{S} w_s(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})$$

and

$$V(\mathbf{w}) = \sum_{s=1}^{S} w_s^2 \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{D}_{\boldsymbol{\theta}} + 2\sum_{s\neq r}\sum w_s w_r \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_r}\mathbf{D}_{\boldsymbol{\theta}}.$$

Theorem 2 shows the asymptotic normality of the averaging estimator with fixed weights, and it implies that $\widehat{\mu}(\mathbf{w})$ is root-n consistent. The asymptotic bias and variance of the averaging estimator are $\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{B}(\mathbf{w})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0$ and $V(\mathbf{w})$, respectively.

From Theorem 2, the AMSE of the averaging estimator $\widehat{\mu}(\mathbf{w})$ is given by

$$A(\mathbf{w}) = \mathbf{w}'\boldsymbol{\Psi}\mathbf{w}, \tag{3.16}$$

where $\boldsymbol{\Psi}$ is an $S \times S$ matrix with the (s,r)th element

$$\Psi_{s,r} = \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{B}_s\boldsymbol{\Pi}_0\boldsymbol{\delta}_0\boldsymbol{\delta}_0'\boldsymbol{\Pi}_0'\mathbf{B}_r' + \mathbf{H}_{\boldsymbol{\Pi}_s}\boldsymbol{\Sigma}\mathbf{H}_{\boldsymbol{\Pi}_r}\right)\mathbf{D}_{\boldsymbol{\theta}}, \tag{3.17}$$

and $\mathbf{B}_s = \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q}$. We then define the optimal fixed-weight vector as

$$\mathbf{w}^{\text{opt}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathbf{w}'\boldsymbol{\Psi}\mathbf{w}, \tag{3.18}$$

which is the value that minimizes the AMSE of $\widehat{\mu}(\mathbf{w})$ over $\mathbf{w} \in \mathcal{W}$. Thus, the averaging estimator with the optimal weights $\widehat{\mu}(\mathbf{w}^{\text{opt}})$ achieves the minimum AMSE in a class of averaging estimators defined by $\widehat{\mu}(\mathbf{w})$.

The optimal weight vector, however, is infeasible, since $\boldsymbol{\Psi}$ is unknown. A feasible version of $\mathbf{w}^{\text{opt}}$ could be obtained by replacing the unknown parameters in $\boldsymbol{\Psi}$ with their sample analogs. As we discussed in Section 3.1, the unknown parameters $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$ can be consistently estimated by the sample analogs. Notice that a consistent estimator for $\boldsymbol{\delta}_0$ is not available

due to the local-to-zero assumption. We therefore follow Wan, Zhang, and Wang (2014) and Liu (2015) and propose a plug-in estimator of $A(\mathbf{w})$ as follows

$$\widehat{A}(\mathbf{w}) = \mathbf{w}'\widehat{\mathbf{\Psi}}\mathbf{w}, \tag{3.19}$$

where the (s,r)th element of $\widehat{\mathbf{\Psi}}$ is

$$\widehat{\Psi}_{s,r} = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\widehat{\mathbf{B}}_s \mathbf{\Pi}_0 \widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} \mathbf{\Pi}_0' \widehat{\mathbf{B}}_r' + \widehat{\mathbf{H}}_{\mathbf{\Pi}_s} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{H}}_{\mathbf{\Pi}_r} \right) \widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \tag{3.20}$$

and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is defined in (3.5). Notice that $\widehat{A}(\mathbf{w})$ is an asymptotically unbiased estimator of $A(\mathbf{w})$.

We now define the plug-in averaging method for extremum estimators. The data-driven weights based on the plug-in method are defined as

$$\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_S)' = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathbf{w}'\widehat{\mathbf{\Psi}}\mathbf{w}. \tag{3.21}$$

When the number of submodels is $S = 2$, we have a closed-form solution to (3.21), and when $S > 2$, the data-driven weights can be found numerically via quadratic programming. We then use $\widehat{\mathbf{w}}$ to construct a plug-in estimator of μ_0 as follows

$$\widehat{\mu}(\widehat{\mathbf{w}}) = \sum_{s=1}^{S} \widehat{w}_s \widehat{\mu}_s. \tag{3.22}$$

As mentioned by Hjort and Claeskens (2003) and Liu (2015), we can also estimate $A(\mathbf{w})$ by inserting $\widehat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}_0$ directly. Thus, the alternative

estimator of $\Psi_{s,r}$ is

$$\widetilde{\Psi}_{s,r} = \widehat{\mathbf{D}}'_{\boldsymbol{\theta}} \left(\widehat{\mathbf{B}}_s \boldsymbol{\Pi}_0 \widehat{\boldsymbol{\delta}} \widehat{\boldsymbol{\delta}}' \boldsymbol{\Pi}'_0 \widehat{\mathbf{B}}'_r + \widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{H}}_{\boldsymbol{\Pi}_r} \right) \widehat{\mathbf{D}}_{\boldsymbol{\theta}}. \tag{3.23}$$

As shown in Section 4, the plug-in averaging method based on (3.23) could have better asymptotic and finite sample properties than the plug-in averaging method based on (3.20).

3.4 Asymptotic behavior of the FIC and plug-in averaging method

In this section, we investigate the limiting distributions of the FIC and the proposed averaging estimator $\widehat{\mu}(\widehat{\mathbf{w}})$. As mentioned in the previous section, $\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$, $\widehat{\mathbf{H}}$, and $\widehat{\boldsymbol{\Sigma}}$ are consistent estimators for $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$, respectively, and $\widehat{\boldsymbol{\delta}} \xrightarrow{d} \mathbf{Z}_{\boldsymbol{\delta}} \sim N(\boldsymbol{\delta}_0, \boldsymbol{\Pi}'_0 \mathbf{H}^{-1} \boldsymbol{\Sigma} \mathbf{H}^{-1} \boldsymbol{\Pi}_0)$ by (3.4). Therefore, it follows that

$$\begin{aligned} \mathrm{FIC}_s \xrightarrow{d} & \mathbf{D}'_{\boldsymbol{\theta}} (\mathbf{H}_{\boldsymbol{\Pi}_s} \mathbf{H} - \mathbf{I}_{p+q}) \boldsymbol{\Pi}_0 (\mathbf{Z}_{\boldsymbol{\delta}} \mathbf{Z}'_{\boldsymbol{\delta}} - \boldsymbol{\Pi}'_0 \mathbf{H}^{-1} \boldsymbol{\Sigma} \mathbf{H}^{-1} \boldsymbol{\Pi}_0) \boldsymbol{\Pi}'_0 (\mathbf{H}_{\boldsymbol{\Pi}_s} \mathbf{H} - \mathbf{I}_{p+q})' \mathbf{D}_{\boldsymbol{\theta}} \\ & + \mathbf{D}'_{\boldsymbol{\theta}} \mathbf{H}_{\boldsymbol{\Pi}_s} \boldsymbol{\Sigma} \mathbf{H}_{\boldsymbol{\Pi}_s} \mathbf{D}_{\boldsymbol{\theta}}. \end{aligned} \tag{3.24}$$

This result shows that the proposed FIC defined in (3.3) will not converge in probability to the AMSE of $\widehat{\mu}_s$, although FIC_s is an asymptotically unbiased estimator of $\mathrm{E}(\Lambda_s^2)$ in (3.2). Furthermore, the above result implies that the FIC model selection estimator has a nonstandard asymptotic distribution. The following corollary presents the limiting distribution of the plug-in estimator $\widehat{\mu}(\widehat{\mathbf{w}})$.

Corollary 2. *Suppose that Assumptions 1–3 hold. Assume that* $\widehat{\mathbf{\Psi}}$ *and* $\mathbf{\Psi}^{\infty}$ *are positive definite. As* $n \to \infty$*, we have*

$$\widehat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^{\infty} = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathbf{w}' \mathbf{\Psi}^{\infty} \mathbf{w} \tag{3.25}$$

and

$$\sqrt{n}(\widehat{\mu}(\widehat{\mathbf{w}}) - \mu_0) \xrightarrow{d} \sum_{s=1}^{S} w_s^{\infty} \Lambda_s, \tag{3.26}$$

where Λ_s *is defined in Corollary 1 and* $\mathbf{\Psi}^{\infty}$ *is an* $S \times S$ *matrix with the* (s, r)*th element*

$$\Psi_{s,r}^{\infty} = \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{B}_s \mathbf{\Pi}_0 (\mathbf{Z}_{\boldsymbol{\delta}} \mathbf{Z}_{\boldsymbol{\delta}}' - \mathbf{\Pi}_0' \mathbf{H}^{-1} \mathbf{\Sigma} \mathbf{H}^{-1} \mathbf{\Pi}_0) \mathbf{\Pi}_0' \mathbf{B}_r' + \mathbf{H}_{\mathbf{\Pi}_s} \mathbf{\Sigma} \mathbf{H}_{\mathbf{\Pi}_r} \right) \mathbf{D}_{\boldsymbol{\theta}}. \tag{3.27}$$

Corollary 2 shows that the data-driven weights (3.21) will not converge in probability to the optimal weights (3.18). Furthermore, the estimated weights are asymptotically random under the local asymptotic framework. This is because the estimate $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is random in the limit. Therefore, unlike the asymptotic normality of the averaging estimator with fixed weights presented in Theorem 2, the averaging estimator with data-driven weights has a nonstandard asymptotic distribution. This non-normal nature of the limiting distribution of the averaging estimator with data-driven weights is pointed out by Hjort and Claeskens (2003) as well as Liu (2015). To address the problem of inference after model averaging, we follow Claeskens

and Carroll (2007), Zhang and Liang (2011), and Liu (2015) to construct a valid confidence interval; see the discussion in the supplementary materials for more details.

Remark 3. Notice that $\mathbf{w}'\boldsymbol{\Psi}^{\infty}\mathbf{w}$ is a convex minimization problem when $\mathbf{w}'\boldsymbol{\Psi}^{\infty}\mathbf{w}$ is quadratic, $\boldsymbol{\Psi}^{\infty}$ is positive definite, and $\mathcal{W}$ is convex. Hence, $\mathbf{w}'\boldsymbol{\Psi}^{\infty}\mathbf{w}$ has a unique minimum; see Charkhi, Claeskens, and Hansen (2016) for more discussion on the uniqueness of the weights. For the estimator defined in (3.23), the estimated weights are still random in the limit since we can show that

$$\widetilde{\Psi}_{s,r} \xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}' \left(\mathbf{B}_s \boldsymbol{\Pi}_0 \mathbf{Z}_{\boldsymbol{\delta}} \mathbf{Z}_{\boldsymbol{\delta}}' \boldsymbol{\Pi}_0' \mathbf{B}_r' + \mathbf{H}_{\boldsymbol{\Pi}_s} \boldsymbol{\Sigma} \mathbf{H}_{\boldsymbol{\Pi}_r}\right) \mathbf{D}_{\boldsymbol{\theta}}. \tag{3.28}$$

Compared to (3.27), the alternative estimator $\widetilde{\Psi}_{s,r}$ has a simpler limiting distribution than the estimator $\widehat{\Psi}_{s,r}$.

Remark 4. Using Theorem 2, we can easily apply the plug-in averaging method to different model setups, and then obtain the asymptotic distribution of the plug-in estimator based on Corollary 2. For example, if $\widehat{Q}_n(\cdot)$ is the sum of squared errors with $h(\mathbf{x}_i, \boldsymbol{\theta}_0) = \mathbf{x}_i'\boldsymbol{\theta}_0$, then Corollary 2 corresponds to Theorem 3 of Liu (2015). Or, if $\widehat{Q}_n(\cdot)$ is the log-likelihood function, then Corollary 2 corresponds to Theorem 1 of Charkhi, Claeskens, and Hansen (2016).

4. Numerical study

In this section, we first evaluate the asymptotic performance of the FIC and plug-in averaging method in a simple three-nested-model framework. We next compare the finite sample performance of the proposed methods with other existing model selection and model averaging methods via Monte Carlo experiments. In the last subsection, we apply the proposed methods to a real data analysis.

4.1 AMSE comparison

We evaluate the asymptotic performance of the different estimates of the focus parameter μ based on the numerical calculation of the AMSE. We consider a simple three-nested-model framework based on the model (3.6), where the model specification is $h(\cdot) = \exp(\mathbf{x}_i'\boldsymbol{\theta})$, $p = 1$, $q = 2$, $M = 3$, $\boldsymbol{\delta}_0 = d(1.5, 1.25)'$, and d varies on a grid between -4 and 4.

We consider the homoskedastic error and set $\sigma^2 = 1$ and $\boldsymbol{\Sigma} = -\sigma^2\mathbf{H}$, where the diagonal elements of $\mathbf{H}$ are -1, and off-diagonal elements are -0.5. The focus parameter is $\mu = \theta_1$, and $\mathbf{D}_{\boldsymbol{\theta}} = (1, 0, 0)'$ in this setting. We compare the AMSE of the following estimators: (1) Narrow model estimator (labeled Narrow); (2) Middle model estimator (labeled Middle); (3) Full model estimator (labeled Full); (4) Averaging estimator with the opti-

mal weights $\mathbf{w}^{\text{opt}}$ defined in (3.6) (labeled W-opt); (5) FIC model selection estimator (labeled FIC); (6) Plug-in averaging method based on (3.20) (labeled PIA-1); and (7) Plug-in averaging method based on (3.23) (labeled PIA-2).

We briefly discuss how to calculate the AMSE for each estimator. The narrow model sets both potentially relevant parameters to zero, that is, $\theta_2 = 0$ and $\theta_3 = 0$. The middle model includes the first potentially relevant parameter and sets the second potentially relevant parameter to zero, while the full model includes both potentially relevant parameters. For these submodel estimators, the AMSE is calculated based on (3.2). For W-opt, we first compute the optimal weights based on (3.18), and then calculate the AMSE by plugging the value of the optimal weights into (3.16). For the FIC, the AMSE is approximated based on (3.24) by simulation averaging across 10,000 random samples. For PIA-1 and PIA-2, the AMSE is approximated based on Corollary 2 by simulation averaging across 10,000 random samples. We divide the AMSE of each estimator by that of W-opt and report the relative AMSE for easy comparison. When the relative AMSE exceeds one, it indicates that the specified estimator has larger AMSE than the averaging estimator with the optimal weights.

Figure 1 presents the relative AMSEs of different estimators. We first

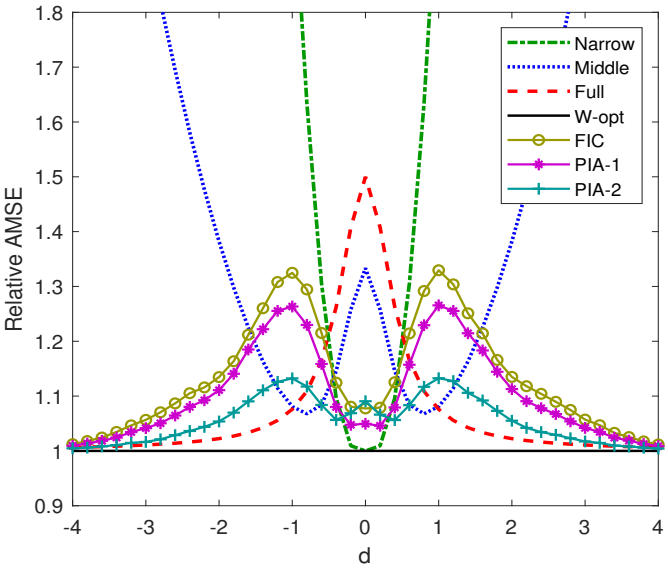

Figure 1: Relative AMSE

compare the AMSEs between the submodel estimators and W-opt. As we expected, the narrow model achieves a lower relative AMSE than the other two submodels for smaller $|d|$, while the full model achieves a smaller relative AMSE than the other two submodels for larger $|d|$. Therefore, the best submodel, which has the lowest AMSE among the submodels, varies with d. Compared to the three submodels, W-opt has much lower AMSEs in most ranges of d. We next compare the AMSEs of FIC, PIA-1, and PIA-2. The numerical results show that PIA-2 has a smaller relative AMSE than PIA-1, and PIA-1 has a smaller relative AMSE than FIC. Notice that the AMSE of PIA-2 is slightly larger than that of W-opt, which illustrates the

effect of the estimated local parameter on asymptotic behavior of the plug-in averaging method. Similarly, for a fixed value of d, the AMSE of FIC is larger than that of the best submodel due to the absence of a consistent estimator for the local parameter. We also compare the model weights of W-opt, PIA-1, and PIA-2 in the supplementary materials.

4.2 Finite sample performance

We next investigate the finite sample performance of the proposed FIC and plug-in averaging methods via Monte Carlo experiments. We consider a nonlinear regression model:

$$y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}) + e_i, \tag{4.1}$$

where $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})' \sim iid$ Uniform$(-1, 1)$ and $\mathbf{z}_i = (z_{1i}, \ldots, z_{qi})' \sim iid$ Uniform$(-1, 1)$. The error term is generated by $e_i = \sigma_i \epsilon_i$, where ϵ_i is generated from a log-normal distribution with mean zero and variance one. For the homoskedastic simulation, we set $\sigma_i = 1$, and for the heteroskedastic simulation, we set $\sigma_i^2 = 0.5 + 1.5x_{pi}^2$. The sample size is $n = 100$ or 250.

We let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ be the must-have parameters and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ the potentially relevant parameters. We set $\beta_j = c$ for $j = 1, \ldots, p$, where the parameter c varies on a grid between -2 and 2, and set $\gamma_k =$

$n^{-1/2}((q-k+1)/q)$ for $k = 1, \ldots, q$. We consider a set of 2^q non-nested submodels and set $p = 1$, 2, or 3, and $q = 3$, 4, or 5. Thus, the numbers of the models are $S = 8$, 16, and 32 for $q = 3$, 4, and 5, respectively.

In addition to FIC, PIA-1, and PIA-2 mentioned in the previous section, we also consider the following estimators: (1) Akaike information criterion model selection estimator (labeled AIC); (2) Bayesian information criterion model selection estimator (labeled BIC); (3) Smoothed AIC model selection estimator (labeled SAIC); and (4) Smoothed BIC model selection estimator (labeled SBIC). Let $\widehat{\sigma}_s^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_{si}^2$, where $\widehat{e}_{si}$ is the NLS residual from the model s. The AIC of the sth model is $\text{AIC}_s = n\log(\widehat{\sigma}_s^2) + 2(p+q_s)$, where $p+q_s$ is the number of parameters in the model s, while the BIC of the sth model is $\text{BIC}_s = n\log(\widehat{\sigma}_s^2) + \log(n)(p+q_s)$. For AIC and BIC, we select the model with the lowest value of AIC or BIC, respectively. The SAIC estimator is proposed by Buckland, Burnham, and Augustin (1997) and it uses the exponential AIC as the model weight. The SAIC weight is proportional to the likelihood of the model and is defined as $\widehat{w}_s = \exp(-\frac{1}{2}\text{AIC}_s)/\sum_{r=1}^S \exp(-\frac{1}{2}\text{AIC}_r)$. The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors, and the SBIC weight is $\widehat{w}_s = \exp(-\frac{1}{2}\text{BIC}_s)/\sum_{r=1}^S \exp(-\frac{1}{2}\text{BIC}_r)$.

Our focus parameter is $\mu = \beta_p$, which is the last element of the must-

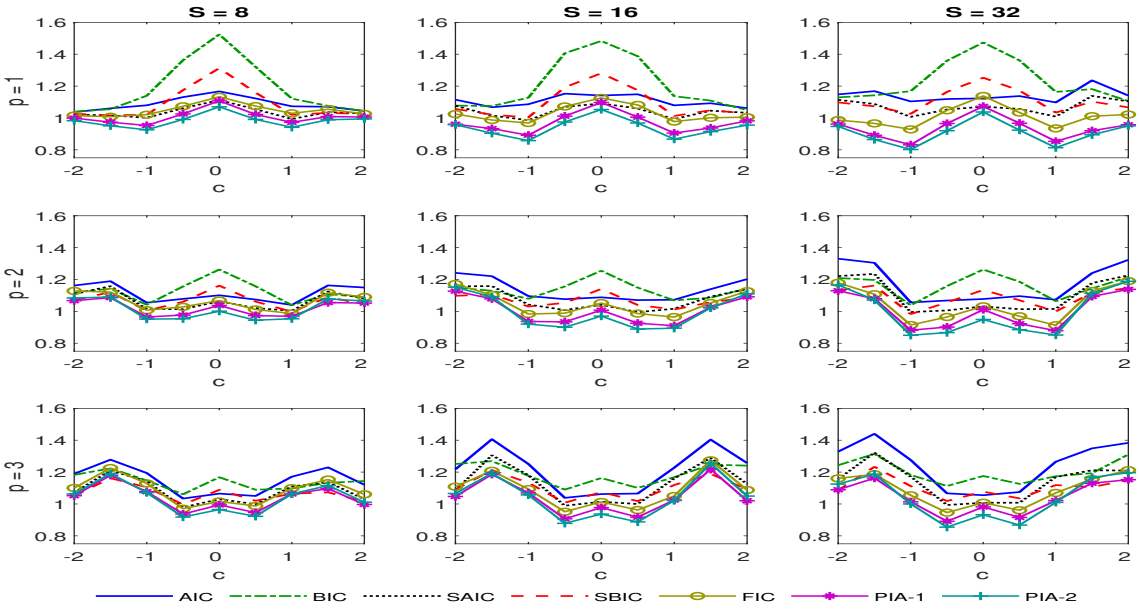

Figure 2: Relative MSE, homoskedastic errors, $n = 100$.

have parameters. To evaluate the finite sample behavior of each estimator, we compare these estimators based on the mean squared error (MSE) of $\widehat{\mu}$. The MSE is calculated by the average of $(\widehat{\mu} - \mu)^2$ obtained from each method over 5,000 replications. For easy comparison, we divide the MSE of each method by that of the best-fitting submodel and report the relative MSE. The best-fitting submodel is the model that has the lowest MSE among all submodels. Therefore, lower relative MSE means better finite sample performance. When the relative MSE exceeds one, it indicates that the specified estimator performs worse than the best-fitting submodel.

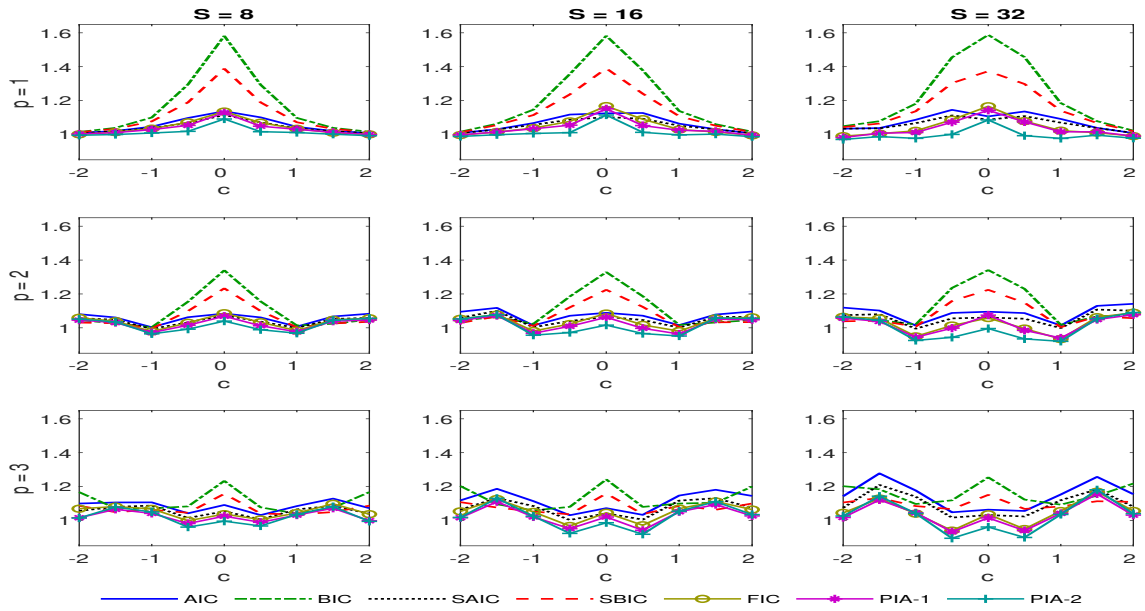

Figure 3: Relative MSE, homoskedastic errors, $n = 250$.

Figures 2 and 3 present the relative MSEs of different estimates in the homoskedastic setup for $n = 100$ and 250, respectively. In each figure, the relative MSEs are displayed for $p = \{1, 2, 3\}$ and $S = \{8, 16, 32\}$ in nine panels, and in each panel, the relative MSEs are displayed for c between -2 and 2. We first compare the finite sample performance of AIC, BIC, SAIC, and SBIC. The simulation results show that BIC has a larger MES than AIC for smaller $|c|$ in all cases, while AIC has a larger MSE than BIC for larger $|c|$ when $p = 2$ and 3. SAIC and SBIC have lower MSEs than AIC and BIC, respectively, and the pattern of relative performance

between SAIC and SBIC is quite similar to that of AIC and BIC. We next compare the finite sample performance of FIC, PIA-1, and PIA-2. The results show that FIC, PIA-1, and PIA-2 perform quite well and have lower MSEs than AIC, BIC, SAIC, and SBIC in most cases. PIA-2 performs slightly better than PIA-1, and PIA-1 performs slightly better than FIC. The relative performance of FIC, PIA-1, and PIA-2 in the finite sample is consistent with our finding in the AMSE comparison presented in Figure 1.

4.3 Real data analysis

In this section, we apply the proposed FIC and plug-in averaging methods to investigate the relationship between income and education. We employ Riphahn, Wambach, and Million (2003)'s German Socioeconomic Panel data set, which is used to study the log-linear model for income in Example 7.6 of Greene (2012). The data consist of 27,326 observations and are available at the Journal of Applied Econometrics data archive website. We follow Greene (2012) and use the last wave of the data set (year 1988) to model incomes. After deleting two observations with zero income, we have a sample of 4,481 observations. The dependent variable is the household monthly net income in German marks, and the explanatory variables include years of schooling (Education), age in years (Age), female (1 = fe-

male, 0 = male), and the quadratic and interaction terms of variables; see Riphahn, Wambach, and Million (2003) for a detailed description of the data.

We follow Greene (2012) and fit an exponential regression model to the data. We assume that the constant term, Education, Age, and Female are must-have regressors, and treat the quadratic and interaction terms of variables as potentially relevant regressors. We consider all possible subsets of potentially relevant regressors, which leads to a total of 32 non-nested models. Our focus parameter is the coefficient of Education. We first estimate the coefficient in each candidate model, and then apply the same model selection and model averaging methods as those in the simulation study.

Table 1 presents the estimation results based on model selection and model averaging methods. The results show that all coefficients have the same signs across different estimation methods except the estimated coefficient of Female by FIC. Furthermore, the coefficient estimates of Education are quite similar across different estimators, while FIC/PIA-1 has a relative larger/smaller coefficient estimate of Education.

We next follow Rolling, Yang, and Velez (2019) and perform a guided simulation experiment to evaluate the different methods under the simula-

Table 1: Estimation results

	AIC	BIC	SAIC	SBIC	FIC	PIA-1	PIA-2
Constant	-3.5731	-3.4776	-3.5534	-3.4840	-2.2245	-3.0197	-3.0962
	(0.2728)	(0.3754)	(0.2824)	(0.3690)	(0.8504)	(0.3533)	(0.4183)
Education	0.1249	0.1217	0.1242	0.1212	0.1279	0.1189	0.1239
	(0.0308)	(0.0452)	(0.0322)	(0.0447)	(0.0384)	(0.0323)	(0.0305)
Age	0.0646	0.0624	0.0642	0.0627	0.0024	0.0408	0.0428
	(0.0067)	(0.0074)	(0.0068)	(0.0073)	(0.0359)	(0.0087)	(0.0120)
Female	0.3941	0.2574	0.3661	0.2720	-0.0024	0.3388	0.3503
	(0.1019)	(0.1428)	(0.1008)	(0.1267)	(0.1510)	(0.0832)	(0.0929)
Education^2	-0.0044	-0.0045	-0.0045	-0.0045	-0.0029	-0.0023	-0.0025
	(0.0011)	(0.0015)	(0.0011)	(0.0015)	(0.0015)	(0.0011)	(0.0011)
Age^2	-0.0009	-0.0009	-0.0009	-0.0009		-0.0004	-0.0004
	(0.0001)	(0.0001)	(0.0001)	(0.0001)		(0.0001)	(0.0001)
Educ × Age	0.0012	0.0013	0.0012	0.0013			
	(0.0003)	(0.0003)	(0.0003)	(0.0003)			
Educ × Female	-0.0224	-0.0212	-0.0221	-0.0211		-0.0206	-0.0210
	(0.0058)	(0.0093)	(0.0061)	(0.0087)		(0.0065)	(0.0059)
Age × Female	-0.0029		-0.0023	-0.0004		-0.0022	-0.0024
	(0.0015)		(0.0014)	(0.0014)		(0.0007)	(0.0013)

*Standard errors, reported in parentheses, are calculated using 1,000 bootstrap replications.

tion scenarios that are consistent with the data. The simulation scenario is based on the submodel selected by AIC, BIC, or FIC. As shown in Table 1, the AIC chooses the full model, the BIC chooses the submodel that excludes the regressor Age × Female, and the FIC chooses the submodel that

only includes the potentially relevant regressor Education[2]. For each model selection method τ, we construct the samples as $y_i^* = \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_\tau + \mathbf{z}_{\tau i}'\widehat{\boldsymbol{\gamma}}_\tau) + e_i^*$, where $\mathbf{z}_{\tau i}'$ are the potentially relevant regressors included in the submodel selected by τ, $\widehat{\boldsymbol{\beta}}_\tau$ and $\widehat{\boldsymbol{\gamma}}_\tau$ are the estimated coefficients from the submodel selected by τ, and e_i^* is an i.i.d. random error. The random error is generated by $e_i^* = \widehat{\sigma}_\tau \epsilon_i$, where $\epsilon_i \sim iid$ Lognormal$(0,1)$ and $\widehat{\sigma}_\tau$ is the standard error estimated from the submodel selected by τ. We then apply the model selection and model averaging methods to the samples $\{y_i^*, \mathbf{x}_i, \mathbf{z}_i\}$ and estimate the focus parameter μ, that is, the coefficient of Education. Notice that the true value of μ is known for each choice of τ. From Table 1, the true values of μ are 0.1249, 0.1217, and 0.1279 for the scenario under AIC, BIC, and FIC, respectively.

Table 2 presents the guided simulation results for three scenarios. We report the bias, the variance (Var), and the MSE of $\widehat{\mu}$ based on 5,000 random draws. The results show that all methods have small negative biases in all scenarios, and model averaging methods achieve lower variances than model selection methods in most scenarios. It is clear that AIC has a lower MSE than BIC, and FIC has a lower MSE than AIC in all scenarios. The MSEs of SAIC are similar to those of AIC, while the MSEs of SBIC are lower than those of BIC. Both PIA-1 and PIA-2 perform quite well and have

Table 2: Guided simulation results

	AIC scenario			BIC scenario			FIC scenario		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
AIC	-0.0690	0.0024	0.0072	-0.0684	0.0024	0.0071	-0.0702	0.0014	0.0063
BIC	-0.1014	0.0022	0.0125	-0.0996	0.0022	0.0121	-0.0947	0.0005	0.0095
SAIC	-0.0731	0.0019	0.0072	-0.0727	0.0019	0.0072	-0.0726	0.0009	0.0062
SBIC	-0.0973	0.0014	0.0109	-0.0960	0.0014	0.0106	-0.0919	0.0003	0.0088
FIC	-0.0686	0.0017	0.0064	-0.0680	0.0017	0.0063	-0.0699	0.0014	0.0062
PIA-1	-0.0703	0.0011	0.0060	-0.0688	0.0010	0.0058	-0.0816	0.0003	0.0070
PIA-2	-0.0639	0.0013	0.0054	-0.0631	0.0013	0.0053	-0.0703	0.0009	0.0058

lower MSEs than other methods in the AIC and BIC scenarios. For the FIC scenario, PIA-2 performs better than PIA-1 and has the lowest MSE among all methods.

5. Conclusion

In this paper, we investigate the limiting distribution of extremum estimators in a local asymptotic framework and propose a focused information criterion and a plug-in averaging method for extremum estimators. We investigate the asymptotic and finite sample properties of the proposed selection and averaging methods. We find that the limiting distributions of the FIC model selection estimator and the averaging estimator with data-driven weights are nonstandard due to the absence of a consistent estimator

for the local parameter. Our numerical results show that the proposed plug-in averaging method achieves lower AMSE and MSE than other methods.

Supplementary Materials

The online supplementary materials include the proofs, additional examples and numerical results, and the details for constructing a valid confidence interval for the post-averaging estimator.

Acknowledgements

We thank the editor, the associate editor, and the two referees for their many constructive comments and suggestions. We also thank the conference participants of Advances in Econometrics 2018, AMES 2019, EcoSta 2019, and ESMA 2019 for their discussions and suggestions. Xinyu Zhang gratefully acknowledges research support from the National Natural Science Foundation of China (71925007, 72091212, 71988101, and 12288201), and the CAS Project for Young Scientists in Basic Research (YSBR-008). Chu-An Liu gratefully acknowledges research support from the Academia Sinica Career Development Award (AS-CDA-110-H02) and the Ministry of Science and Technology of Taiwan (MOST 107-2410-H-001-031-MY3). All errors and omissions are our own responsibility.

References

Behl, P., G. Claeskens, and H. Dette (2014). Focused model selection in quantile regression. *Statistica Sinica 24*, 601–624.

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

Chang, M. and F. J. DiTraglia (2018). A generalized focused information criterion for GMM. *Journal of Applied Econometrics 33*(3), 378–397.

Charkhi, A., G. Claeskens, and B. E. Hansen (2016). Minimum mean squared error model averaging in likelihood models. *Statistica Sinica 26*, 809–840.

Chen, X., G. Zou, and X. Zhang (2013). Frequentist model averaging for linear mixed-effects models. *Frontiers of Mathematics in China 8*(3), 497–515.

Cheng, T.-C. F., C.-K. Ing, and S.-H. Yu (2015). Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics 189*(2), 321–334.

Claeskens, G. and R. J. Carroll (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika 94*, 249–265.

Claeskens, G., C. Croux, and J. Van Kerckhoven (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics 62*(4), 972–979.

Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association 98*(464), 900–916.

Claeskens, G. and N. L. Hjort (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Greene, W. H. (2012). *Econometric Analysis*. Seventh Edition. Pearson.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory 21*(21), 60–68.

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

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the*

American Statistical Association 98, 879–899.

Hjort, N. L. and G. Claeskens (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association 101*(476), 1449–1464.

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

Jullum, M. and N. L. Hjort (2017). Parametric or nonparametric: the fic approach. *Statistica Sinica 27*, 951–981.

Kitagawa, T. and C. Muris (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics 193*, 271–289.

Leeb, H. and B. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*, 142–159.

Lohmeyer, J., F. Palm, H. Reuvers, and J.-P. Urbain (2019). Focused information criterion for locally misspecified vector autoregressive models. *Econometric Reviews 38*(7), 763–792.

Lu, X. (2015). A covariate selection criterion for estimation of treatment effects. *Journal of Business and Economic Statistics 33*, 506–522.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys 29*(1), 46–75.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier.

Pircalabelu, E., G. Claeskens, and L. Waldorp (2015). A focused information criterion for graphical models. *Statistics and Computing 25*, 1071–1092.

Riphahn, R. T., A. Wambach, and A. Million (2003). Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics 18*(4), 387–405.

Rolling, C. A., Y. Yang, and D. Velez (2019). Combining estimates of conditional treatment effects. *Econometric Theory 35*(6), 1089–1110.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Litera-*

ture 58(3), 644–719.

Sueishi, N. (2013). Generalized empirical likelihood-based focused information criterion and model averaging. *Econometrics 1*, 141–156.

Wan, A. T., X. Zhang, and S. Wang (2014). Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting 30*(1), 118–128.

Wang, H. and C. Leng (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association 102*(479), 1039–1048.

Wang, H., G. Zou, and A. T. K. Wan (2012). Model averaging for varying-coefficient partially linear measurement error models. *Electronic Journal of Statistics 6*, 1017–1039.

Xu, G., S. Wang, and J. Z. Huang (2014). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics 41*(2), 365–381.

Yu, Y., S. W. Thurston, R. Hauser, and H. Liang (2013). Model averaging procedure for partially linear single-index models. *Journal of Statistical Planning and Inference 143*, 2160–2170.

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

Zhang, X., A. T. K. Wan, and S. Z. Zhou (2012). Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *Journal of Business and Economic Statistics 30*, 132–142.

Xinyu Zhang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

E-mail: xinyu@amss.ac.cn

Chu-An Liu

Institute of Economics, Academia Sinica

E-mail: caliu@econ.sinica.edu.tw

A UNIFIED APPROACH TO FOCUSED INFORMATION CRITERION AND PLUG-IN AVERAGING METHOD

Xinyu Zhang[1,2] and Chu-An Liu[3]

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

[2]*Beijing Academy of Artificial Intelligence*

[3]*Institute of Economics, Academia Sinica*

Supplementary Material

The online supplementary materials include five parts. S1 contains the proofs of theorems and corollaries. S2 verifies the high-level assumptions for the nonlinear least squares estimator example. S3 provides additional examples to illustrate the general results from Section 3.1. S4 provides the model weights of W-opt, PIA-1, and PIA-2 in a simple three-nested-model framework and additional simulation results for the heteroskedastic setup. S5 describes the details for constructing a valid confidence interval for the post-averaging estimator.

S1. Proofs

Proof of Theorem 1: The proof has two steps. In the first step, we prove $\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2})$. In the second step, we show that $\widehat{\boldsymbol{\theta}}_s$ is approximatively a linear function of $\widehat{\boldsymbol{\theta}}$, by which we get the conclusion of Theorem 1.

Step 1. By Assumptions 1 and 3 and Equation (2.3), we have

$$\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2}). \tag{S1.1}$$

For any $\boldsymbol{\theta} \in \mathbb{R}^{p+q}$, by a Taylor expansion and the fact that $\partial\widehat{Q}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$, we have

$$\begin{aligned}\widehat{Q}_n(\boldsymbol{\theta}) = \widehat{Q}_n(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{H}_n(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ - \frac{1}{6}\sum_{i=1}^{p+q}\sum_{j=1}^{p+q}\sum_{k=1}^{p+q}\left(\frac{\partial^3\widehat{Q}_n(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)(\widehat{\theta}_i - \theta_i)(\widehat{\theta}_j - \theta_j)(\widehat{\theta}_k - \theta_k),\end{aligned} \tag{S1.2}$$

where $\boldsymbol{\theta}^*$ is a vector between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. We then follow the argument in Fan and Li (2001) and Wang and Leng (2007) to show that for any given $\epsilon > 0$, there exists a large positive constant C_s such that

$$\liminf_n \Pr\left\{\inf_{\|\mathbf{u}_s\|=C_s}\widehat{Q}_n\left(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} + n^{-1/2}\mathbf{u}_s\right) < \widehat{Q}_n\left(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}}\right)\right\} > 1 - \epsilon, \tag{S1.3}$$

where $\mathbf{u}_s$ is a $(p+q)$-dimensional vector with $\boldsymbol{\Pi}_{s^c}\mathbf{u}_s = \mathbf{0}$ and $\|\mathbf{u}_s\| = C_s$.

By (S1.1), (S1.2), and Assumption 2, it follows that

$$\widehat{Q}_n(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}}) = \widehat{Q}_n(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}})'\mathbf{H}_n(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}}) + o_p(n^{-1}). \tag{S1.4}$$

Using (S1.1) and the fact that $\|\mathbf{u}_s\| = C_s$, we have

$$\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} + n^{-1/2}\mathbf{u}_s - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2}). \tag{S1.5}$$

Then by (S1.2), (S1.5), and Assumption 2, it follows that

$$\begin{aligned}\widehat{Q}_n\left(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} + n^{-1/2}\mathbf{u}_s\right) = \widehat{Q}_n(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} - n^{-1/2}\mathbf{u}_s\right)' \\ \times \mathbf{H}_n(\widehat{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} - n^{-1/2}\mathbf{u}_s\right) + o_p(n^{-1}).\end{aligned} \tag{S1.6}$$

Subtracting (S1.4) from (S1.6), we have

$$\begin{aligned}&\widehat{Q}_n\left(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} + n^{-1/2}\mathbf{u}_s\right) - \widehat{Q}_n\left(\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}}\right) \\ &\quad = \frac{1}{n}\left\{\frac{1}{2}\mathbf{u}_s'\mathbf{H}_n(\widehat{\boldsymbol{\theta}})\mathbf{u}_s - \mathbf{u}_s'\mathbf{H}_n(\widehat{\boldsymbol{\theta}})\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}}) + o_p(1)\right\},\end{aligned}$$

which together with (S1.1) and Assumption 1(v) implies (S1.3). Hence, with probability at least $1-\epsilon$, the maximizer $\widehat{\boldsymbol{\theta}}_s$ of $\widehat{Q}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_s, \mathbf{0})$ is in the ball $\left\{\boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} + n^{-1/2}\mathbf{u}_s : \boldsymbol{\Pi}_{s^c}\mathbf{u}_s = \mathbf{0},\ \|\mathbf{u}_s\| = C_s\right\}$. Therefore, we have

$$\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\Pi}_s'\boldsymbol{\Pi}_s\widehat{\boldsymbol{\theta}} = O_p(n^{-1/2}). \tag{S1.7}$$

From (S1.1) and (S1.7), it follows that

$$\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2}). \tag{S1.8}$$

Step 2. Recall that $\boldsymbol{\eta}_s = (\boldsymbol{\beta}', \boldsymbol{\gamma}_s')'$ and $\widehat{\boldsymbol{\theta}}_s = \boldsymbol{\Pi}_s'\widehat{\boldsymbol{\eta}}_s$. Let $\widetilde{Q}_n(\boldsymbol{\eta}_s) \equiv \widehat{Q}_n(\boldsymbol{\Pi}_s'\boldsymbol{\eta}_s) = \widehat{Q}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_s, \mathbf{0})$. By a Taylor expansion, Equation (S1.8), and

Assumption 2, it follows that

$$\begin{aligned}\frac{\partial \widetilde{Q}_n(\boldsymbol{\eta}_s)}{\partial \boldsymbol{\eta}_s}|_{\boldsymbol{\eta}_s=\widehat{\boldsymbol{\eta}}_s} &= \frac{\partial \widehat{Q}_n(\boldsymbol{\Pi}_s'\boldsymbol{\eta}_s)}{\partial \boldsymbol{\eta}_s}|_{\boldsymbol{\eta}_s=\widehat{\boldsymbol{\eta}}_s} = \boldsymbol{\Pi}_s \frac{\partial \widehat{Q}_n(\boldsymbol{\Pi}_s'\boldsymbol{\eta}_s)}{\partial (\boldsymbol{\Pi}_s'\boldsymbol{\eta}_s)}|_{\boldsymbol{\Pi}_s'\boldsymbol{\eta}_s=\boldsymbol{\Pi}_s'\widehat{\boldsymbol{\eta}}_s} \\ &= \boldsymbol{\Pi}_s \frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_s} \\ &= \boldsymbol{\Pi}_s \left\{ \frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \mathbf{H}_n(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}_s'\widehat{\boldsymbol{\eta}}_s) + o_p(n^{-1/2}) \right\}. \end{aligned} \tag{S1.9}$$

By inserting $\partial \widetilde{Q}_n(\boldsymbol{\eta}_s)/\partial \boldsymbol{\eta}_s|_{\boldsymbol{\eta}_s=\widehat{\boldsymbol{\eta}}_s} = \mathbf{0}$ and $\partial \widehat{Q}_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$ into (S1.9), we have

$$\widehat{\boldsymbol{\theta}}_s = \boldsymbol{\Pi}_s'\widehat{\boldsymbol{\eta}}_s = \boldsymbol{\Pi}_s' \left(\boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \boldsymbol{\Pi}_s' \right)^{-1} \boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \widehat{\boldsymbol{\theta}} + o_p(n^{-1/2}). \tag{S1.10}$$

Therefore, we have

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) &= \boldsymbol{\Pi}_s' \left(\boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \boldsymbol{\Pi}_s' \right)^{-1} \boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \sqrt{n}\widehat{\boldsymbol{\theta}} - \sqrt{n}\boldsymbol{\theta}_0^* + o_p(1) \\ &= \boldsymbol{\Pi}_s' \left(\boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \boldsymbol{\Pi}_s' \right)^{-1} \boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\quad + \boldsymbol{\Pi}_s' \left(\boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \boldsymbol{\Pi}_s' \right)^{-1} \boldsymbol{\Pi}_s \mathbf{H}_n(\widehat{\boldsymbol{\theta}}) \sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^*) + o_p(1). \end{aligned} \tag{S1.11}$$

Recall $\boldsymbol{\Pi}_0 = (\mathbf{0}_{q\times p}, \mathbf{I}_q)'$. By Assumption 3, we have $\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^*) = (\mathbf{0}', \boldsymbol{\delta}_0')' = \boldsymbol{\Pi}_0\boldsymbol{\delta}_0$. Then by (2.3), (S1.11), and Assumption 1, we can obtain (2.4). This completes the proof. ■

Proof of Corollary 1: By a Taylor expansion of $\mu(\boldsymbol{\theta}_0)$ and $\mu(\widehat{\boldsymbol{\theta}}_s)$ about $\boldsymbol{\theta}_0^*$, it follows that

$$\mu_0 = \mu(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}_0^*) + \mathbf{D}_{\boldsymbol{\gamma}}'\boldsymbol{\delta}_0/\sqrt{n} + o(n^{-1/2}),$$

$$\widehat{\mu}_s = \mu(\widehat{\boldsymbol{\theta}}_s) = \mu(\boldsymbol{\theta}_0^*) + \mathbf{D}_{\boldsymbol{\theta}}'(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) + o(n^{-1/2}).$$

By the above two equations, Theorem 1 and the application of the delta method, we have

$$\begin{aligned}\sqrt{n}(\widehat{\mu}_s - \mu_0) &= \mathbf{D}_{\boldsymbol{\theta}}'\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) - \mathbf{D}_{\boldsymbol{\gamma}}'\boldsymbol{\delta}_0 + o_p(1) \\ &\xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z} - \mathbf{D}_{\boldsymbol{\theta}}'\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 \\ &= \mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z}.\end{aligned}$$

This completes the proof. ■

Proof of Theorem 2: From Corollary 1, we observe that all of Λ_s can be expressed in terms of the same normal vector $\mathbf{Z}$. Therefore, there is joint convergence in distribution of all $\sqrt{n}(\widehat{\mu}_s - \mu_0)$ to Λ_s for $s = 1, \ldots, S$. Next, notice that the weights are nonrandom. Then, it follows that

$$\sqrt{n}(\widehat{\mu}(\mathbf{w}) - \mu_0) = \sum_{s=1}^{S} w_s\sqrt{n}(\widehat{\mu}_s - \mu_0) \xrightarrow{d} \sum_{s=1}^{S} w_s\Lambda_s \equiv \Lambda(\mathbf{w}).$$

Thus, the asymptotic distribution of the averaging estimator is a weighted average of the normal distributions, which is also a normal distribution.

By standard algebra, we can show the mean of $\Lambda(\mathbf{w})$ as

$$\mathrm{E}\left(\sum_{s=1}^{S} w_s\Lambda_s\right) = \sum_{s=1}^{S} w_s\mathrm{E}(\Lambda_s) = \sum_{s=1}^{S} w_s\mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 = \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{B}(\mathbf{w})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0,$$

where $\mathbf{B}(\mathbf{w}) = \sum_{s=1}^{S} w_s(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})$. We next show the covariance

matrix of $\Lambda(\mathbf{w})$. Let $\mathbf{B}_s = \mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q}$. Then we can rewrite Λ_s as $\Lambda_s = \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{B}_s\mathbf{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}\mathbf{Z}$. For any two submodels, we have

$$\begin{aligned} Cov(\Lambda_s, \Lambda_r) &= \mathrm{E}\left((\mathbf{D}'_{\boldsymbol{\theta}}\mathbf{B}_s\mathbf{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}\mathbf{Z} - \mathrm{E}(\mathbf{D}'_{\boldsymbol{\theta}}\mathbf{B}_s\mathbf{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}\mathbf{Z})) \right. \\ &\qquad \left. \times (\mathbf{D}'_{\boldsymbol{\theta}}\mathbf{B}_r\mathbf{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_r}\mathbf{H}\mathbf{Z} - \mathrm{E}(\mathbf{D}'_{\boldsymbol{\theta}}\mathbf{B}_r\mathbf{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_r}\mathbf{H}\mathbf{Z}))'\right) \\ &= \mathrm{E}(\mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}\mathbf{Z}\mathbf{Z}'\mathbf{H}\mathbf{H}_{\mathbf{\Pi}_r}\mathbf{D}_{\boldsymbol{\theta}}) \\ &= \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{\Sigma}\mathbf{H}_{\mathbf{\Pi}_r}\mathbf{D}_{\boldsymbol{\theta}}. \end{aligned}$$

Therefore, the covariance matrix of $\Lambda(\mathbf{w})$ is

$$\begin{aligned} Var\left(\sum_{s=1}^{S} w_s\Lambda_s\right) &= \sum_{s=1}^{S} w_s^2 Var(\Lambda_s) + 2\sum_{s\neq r}\sum w_s w_r Cov(\Lambda_s, \Lambda_r) \\ &= \sum_{s=1}^{S} w_s^2 \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{\Sigma}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{D}_{\boldsymbol{\theta}} + 2\sum_{s\neq r}\sum w_s w_r \mathbf{D}'_{\boldsymbol{\theta}}\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{\Sigma}\mathbf{H}_{\mathbf{\Pi}_r}\mathbf{D}_{\boldsymbol{\theta}}. \end{aligned}$$

This completes the proof. ■

Proof of Corollary 2: We first show the limiting distribution of $\widehat{\mathbf{w}}$. By Theorem 1, we have $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0^*$, which implies that $\widehat{\mathbf{D}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{D}_{\boldsymbol{\theta}}$. Next, by Theorem 4.1 of Newey and McFadden (1994), we have $\widehat{\mathbf{H}} \xrightarrow{p} \mathbf{H}$ and $\widehat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}$. Recall that $\widehat{\boldsymbol{\delta}} \xrightarrow{d} \mathbf{Z}_{\boldsymbol{\delta}} = \boldsymbol{\delta}_0 + \mathbf{\Pi}_0'\mathbf{Z}$, where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1})$. Then, by the continuous mapping theorem and Slutsky's theorem, it follows that $\widehat{\Psi}_{s,r} \xrightarrow{d} \Psi_{s,r}^{\infty}$. Since all of $\Psi_{s,r}^{\infty}$ can be expressed in terms of the same normal vector $\mathbf{Z}$, there is joint convergence in distribution of all $\widehat{\Psi}_{s,r}$ to $\Psi_{s,r}^{\infty}$. Hence, it follows that $\mathbf{w}'\widehat{\mathbf{\Psi}}\mathbf{w} \xrightarrow{d} \mathbf{w}'\mathbf{\Psi}^{\infty}\mathbf{w}$. Note that $\mathbf{w}'\mathbf{\Psi}^{\infty}\mathbf{w}$ is a

convex minimization problem when $\mathbf{w}'\mathbf{\Psi}^{\infty}\mathbf{w}$ is quadratic, $\mathbf{\Psi}^{\infty}$ is positive definite, and $\mathcal{W}$ is convex. Hence, the limiting process has a unique minimum. Therefore, by Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), the minimizer $\widehat{\mathbf{w}}$ converges in distribution to the minimizer of $\mathbf{w}'\mathbf{\Psi}^{\infty}\mathbf{w}$, which is $\mathbf{w}^{\infty}$.

We next show the asymptotic distribution of $\widehat{\mu}(\widehat{\mathbf{w}})$. Observe that there is joint convergence in distribution of all $\widehat{\mu}_s$ and $\widehat{w}_s$, since both Λ_s and $\mathbf{w}^{\infty}$ can be expressed in terms of the same normal vector $\mathbf{Z}$. Therefore, it follows that

$$\sqrt{n}(\widehat{\mu}(\widehat{\mathbf{w}}) - \mu_0) = \sum_{s=1}^{S} \widehat{w}_s \sqrt{n}(\widehat{\mu}_s - \mu_0) \xrightarrow{d} \sum_{s=1}^{S} w_s^{\infty} \Lambda_s.$$

This completes the proof. ■

S2. Verifications of Assumptions in the nonlinear least squares estimator example.

We now verify the high-level assumptions for the nonlinear least squares estimator in Section 3.2. Let $S(\boldsymbol{\theta}) = \mathrm{E}((y_i - h(\mathbf{x}_i, \boldsymbol{\theta}))^2)$. For Assumption 1(i), the primitive conditions are $\mathrm{E}(y_i^2) < \infty$, $\mathrm{E}|h(\mathbf{x}_i, \boldsymbol{\theta}_0)|^2 < \infty$, and $S(\boldsymbol{\theta}) > S(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, and for Assumption 1(iv), a simple sufficient condition is $\mathrm{E}(y_i^4) < \infty$, $\mathrm{E}|h(\mathbf{x}_i, \boldsymbol{\theta}_0)|^4 < \infty$, $\mathrm{E}\|\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}_i, \boldsymbol{\theta})\|^4 < \infty$, and $\mathrm{E}\|\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} h(\mathbf{x}_i, \boldsymbol{\theta})\|^4 < \infty$; see p.777-778 of Hansen (2022) for a detailed

discussion.

We next provide the primitive assumptions for Assumption 2. We can show that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_0^* \in \boldsymbol{\Theta}_0^*} \frac{\partial^3 \widehat{Q}_n(\boldsymbol{\theta})}{\partial\theta_l \partial\theta_j \partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \\
&= \sup_{\boldsymbol{\theta}_0^* \in \boldsymbol{\Theta}_0^*} \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\partial^2 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_l \partial\theta_j} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} + \frac{\partial^2 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_l \partial\theta_k} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_j}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right. \\
&\quad \left. + \frac{\partial^2 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_i}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} - e_i \frac{\partial^3 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_l \partial\theta_j \partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right\}. \qquad \text{(S2.1)}
\end{aligned}$$

Therefore, Assumption 2 holds in this example if

$$\sup_{\boldsymbol{\theta}_0^* \in \boldsymbol{\Theta}_0^*} \left| \frac{\partial^2 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_l}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right| = o_p(n^{1/2}) \qquad \text{(S2.2)}$$

and

$$\sup_{\boldsymbol{\theta}_0^* \in \boldsymbol{\Theta}_0^*} \left| e_i \frac{\partial^3 h(\mathbf{x}_i, \boldsymbol{\theta})}{\partial\theta_l \partial\theta_j \partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*} \right| = o_p(n^{1/2}) \qquad \text{(S2.3)}$$

for $l, j, k \in \{1, \ldots, p+q\}$. Note that these two conditions imply that we allow the left-hand side of Equations (S2.2) and (S2.3) to diverge with the sample size at a rate slower than $n^{1/2}$.

S3. Additional examples

In this section, we provide additional examples to illustrate the general results from Section 3.1. Examples include the maximum likelihood estimator (MLE), the generalized method of moments (GMM) estimator, and the minimum distance (MD) estimator.

S3.1 Maximum likelihood estimator

Suppose the data $(z_1, \ldots, z_n)$ are i.i.d. with the density function $f(z|\boldsymbol{\theta}_0)$ and unknown parameters $\boldsymbol{\theta}_0$. The likelihood function is $\Pi_{i=1}^n f(z_i|\boldsymbol{\theta}_0)$ and the log-likelihood function is $\sum_{i=1}^n \log f(z_i|\boldsymbol{\theta}_0)$. The MLE estimator $\widehat{\boldsymbol{\theta}}$ maximizes the log-likelihood function

$$\widehat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(z_i|\boldsymbol{\theta}). \tag{S3.1}$$

Note that the objective function $\widehat{Q}_n(\boldsymbol{\theta})$ converges to $Q_0(\boldsymbol{\theta}) = \mathrm{E}(\log f(z_i|\boldsymbol{\theta}))$. Thus,

$$\mathbf{H}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(z_i|\boldsymbol{\theta}), \quad \mathbf{H}(\boldsymbol{\theta}) = E\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(z_i|\boldsymbol{\theta}), \tag{S3.2}$$

and

$$\boldsymbol{\Sigma} = \mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(z_i|\boldsymbol{\theta}_0)\frac{\partial}{\partial\boldsymbol{\theta}'}\log f(z_i|\boldsymbol{\theta}_0)\right) \equiv \mathbf{J}, \tag{S3.3}$$

where $\mathbf{J}$ is called the information matrix. When the information matrix equality holds, we have $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}_0) = -\mathbf{J}$. This result together with Theorem 1 shows that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) \xrightarrow{d} \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{J}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s\mathbf{J}(\mathbf{Z} + \boldsymbol{\Pi}_0\boldsymbol{\delta}_0) \sim N\left(\mathbf{J}_{\boldsymbol{\Pi}_s}\mathbf{J}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0, \mathbf{J}_{\boldsymbol{\Pi}_s}\right), \tag{S3.4}$$

where $\mathbf{J}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{J}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$. By Corollary 1 and some algebra, we have

$$\mathrm{E}(\Lambda_s^2) = \mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{J}_{\boldsymbol{\Pi}_s}\mathbf{J} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0\boldsymbol{\delta}_0'\boldsymbol{\Pi}_0'(\mathbf{J}_{\boldsymbol{\Pi}_s}\mathbf{J} - \mathbf{I}_{p+q})'\mathbf{D}_{\boldsymbol{\theta}} + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{J}_{\boldsymbol{\Pi}_s}\mathbf{D}_{\boldsymbol{\theta}}. \tag{S3.5}$$

Thus, the FIC for the MLE estimator is defined as

$$\mathrm{FIC}_s = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'(\widehat{\mathbf{J}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{J}} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\boldsymbol{\Pi}_0'(\widehat{\mathbf{J}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{J}} - \mathbf{I}_{p+q})'\widehat{\mathbf{D}}_{\boldsymbol{\theta}} + \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{J}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \quad \text{(S3.6)}$$

where $\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$ and $\widehat{\mathbf{J}}$ are the sample analogs of $\mathbf{D}_{\boldsymbol{\theta}}$ and $\mathbf{J}$, and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'} = \widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}' - \boldsymbol{\Pi}_0'\widehat{\mathbf{J}}^{-1}\boldsymbol{\Pi}_0$ is the asymptotically unbiased estimator of $\boldsymbol{\delta}_0\boldsymbol{\delta}_0'$.

Remark 1. Hjort and Claeskens (2003) and Claeskens and Hjort (2003) investigate the limiting distribution of the MLE estimator in a local asymptotic framework and develop FIC under the likelihood framework. Our result (S3.4) corresponds to Lemma 3.2 of Hjort and Claeskens (2003), and the FIC given in (S3.6) corresponds to the equation (3.3) in Claeskens and Hjort (2003).

Remark 2. Using Theorem 1, we can easily obtain the asymptotic normality of the submodel estimator and construct the FIC for different likelihood model setups. For example, if $\widehat{Q}_n(\cdot)$ is the log-partial likelihood as in the equation (3) of Hjort and Claeskens (2006), we can obtain their Lemma 1 and construct the FIC for the Cox hazard regression model. Or, if $\widehat{Q}_n(\cdot)$ is the quasi-likelihood function as in the equation (2.2) of Zhang and Liang (2011), we can obtain their Theorem 1 and construct the FIC for generalized additive partial linear models.

S3.2 Generalized method of moments estimator

Let $g(\mathbf{z}, \boldsymbol{\theta})$ be an $\ell \times 1$ vector of moment functions and $\boldsymbol{\theta}$ a $k \times 1$ vector of unknown parameters with $\ell \geq k$. Suppose the data $\mathbf{z}_i$ are i.i.d. and the moment conditions satisfy $\mathrm{E}(g(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$. Let $\mathbf{W}_n$ be an $\ell \times \ell$ positive semi-definite weight matrix. The GMM estimator $\widehat{\boldsymbol{\theta}}$ maximizes the following objective function

$$\widehat{Q}_n(\boldsymbol{\theta}) = -\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i, \boldsymbol{\theta})\right)' \mathbf{W}_n \left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i, \boldsymbol{\theta})\right). \tag{S3.7}$$

Note that the GMM estimator includes the linear instrumental variable estimator as a special case when $g(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{x}_i(y_i - \mathbf{Y}_i'\boldsymbol{\theta})$, where y_i is a dependent variable, $\mathbf{Y}_i$ are endogenous variables, and $\mathbf{x}_i$ are instrumental variables.

Suppose that $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ and $\widehat{g}_n \equiv \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i, \boldsymbol{\theta}) \xrightarrow{p} \mathrm{E}(g(\mathbf{z}_i, \boldsymbol{\theta})) \equiv g_0(\boldsymbol{\theta})$. Then the objective function $\widehat{Q}_n(\boldsymbol{\theta})$ converges to $Q_0(\boldsymbol{\theta}) = -g_0(\boldsymbol{\theta})'\mathbf{W}g_0(\boldsymbol{\theta})$. Let $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_0) = \mathrm{E}(\frac{\partial}{\partial \boldsymbol{\theta}'} g(\mathbf{z}_i, \boldsymbol{\theta}_0))$ and $\boldsymbol{\Omega} = \mathrm{E}(g(\mathbf{z}_i, \boldsymbol{\theta}_0)g(\mathbf{z}_i, \boldsymbol{\theta}_0)')$. By Assumption 1 and some algebra, we have $\mathbf{H} = -\mathbf{G}'\mathbf{W}\mathbf{G}$ and $\boldsymbol{\Sigma} = \mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}$.

By Theorem 1, it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_0^*) \xrightarrow{d} \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{G}'\mathbf{W}\mathbf{G}(\mathbf{Z} + \boldsymbol{\Pi}_0\boldsymbol{\delta}_0) \sim N(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{G}'\mathbf{W}\mathbf{G}\boldsymbol{\Pi}_0\boldsymbol{\delta}_0, \mathbf{V}_{\boldsymbol{\Pi}_s}), \tag{S3.8}$$

where $\mathbf{H}_{\boldsymbol{\Pi}_s} = -\boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{G}'\mathbf{W}\mathbf{G}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$ and $\mathbf{V}_{\boldsymbol{\Pi}_s} = \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}\mathbf{H}_{\boldsymbol{\Pi}_s}$.

Thus, by Corollary 1, the FIC for the GMM estimator is defined as

$$\begin{aligned}\mathrm{FIC}_s = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{G}}'\mathbf{W}_n\widehat{\mathbf{G}} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\boldsymbol{\Pi}_0'(\widehat{\mathbf{H}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{G}}'\mathbf{W}_n\widehat{\mathbf{G}} - \mathbf{I}_{p+q})'\widehat{\mathbf{D}}_{\boldsymbol{\theta}} \\ + \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{V}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \qquad \text{(S3.9)}\end{aligned}$$

where $\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$, $\widehat{\mathbf{G}}$, and $\widehat{\boldsymbol{\Omega}}$ are the sample analogs of $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{G}$, and $\boldsymbol{\Omega}$, and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is the asymptotically unbiased estimator of $\boldsymbol{\delta}_0\boldsymbol{\delta}_0'$.

For the efficient GMM estimator, we set the weight matrix as $\mathbf{W} = \boldsymbol{\Omega}^{-1}$. Then it follows that $-\mathbf{H} = \boldsymbol{\Sigma} = \mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G} \equiv \mathbf{V}$, and the covariance matrix in (S3.8) is simplified as $\mathbf{V}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{V}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$. In this case, the FIC for the efficient GMM estimator is defined as

$$\begin{aligned}\mathrm{FIC}_s = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'(\widehat{\mathbf{V}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{V}} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}\boldsymbol{\Pi}_0'(\widehat{\mathbf{V}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{V}} - \mathbf{I}_{p+q})'\widehat{\mathbf{D}}_{\boldsymbol{\theta}} \\ + \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{V}}_{\boldsymbol{\Pi}_s}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}, \qquad \text{(S3.10)}\end{aligned}$$

where $\widehat{\mathbf{V}}$ is the sample analog of $\mathbf{V}$.

Remark 3. DiTraglia (2016) and Chang and DiTraglia (2018) propose a focused moment selection criterion for the GMM estimator with a set of locally misspecified moment conditions, i.e., $\mathrm{E}(g(\mathbf{z}, \boldsymbol{\theta}_0)) = n^{-1/2}\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is an unknown constant vector. Although we have focused on the case where the moment conditions are correct, i.e., $\mathrm{E}(g(\mathbf{z}, \boldsymbol{\theta}_0)) = \mathbf{0}$, our results can be easily extended to the case considered in DiTraglia (2016) and Chang and DiTraglia (2018).

S3.3 Minimum distance estimator

Let $h(\boldsymbol{\theta})$ be a function that maps from a $k\times 1$ vector of structural parameters $\boldsymbol{\theta}$ to an $\ell \times 1$ vector of reduced form parameters $\boldsymbol{\alpha}$, where $\ell \geq k$. Suppose that $\widehat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_0 = h(\boldsymbol{\theta}_0)$. Let $\mathbf{W}_n$ be an $\ell \times \ell$ positive semi-definite weight matrix. The MD estimator $\widehat{\boldsymbol{\theta}}$ maximizes the following objective function

$$\widehat{Q}_n(\boldsymbol{\theta}) = -(\widehat{\boldsymbol{\alpha}} - h(\boldsymbol{\theta}))'\mathbf{W}_n(\widehat{\boldsymbol{\alpha}} - h(\boldsymbol{\theta})). \tag{S3.11}$$

Suppose that $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ and $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$. Then the objective function $\widehat{Q}_n(\boldsymbol{\theta})$ converges to $Q_0(\boldsymbol{\theta}) = -(\boldsymbol{\alpha}_0 - h(\boldsymbol{\theta}))'\mathbf{W}(\boldsymbol{\alpha}_0 - h(\boldsymbol{\theta}))$. Let $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta}_0) = \mathrm{E}(\frac{\partial}{\partial \boldsymbol{\theta}'} h(\boldsymbol{\theta}_0))$. By some algebra, we have $\mathbf{H} = -\mathbf{G}'\mathbf{W}\mathbf{G}$ and $\boldsymbol{\Sigma} = \mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}$, where $\mathbf{H}$ and $\boldsymbol{\Sigma}$ have the same sandwich form as those of the GMM estimator. Thus, the FIC for the MD estimator has the same form as (S3.9).

Similar to the GMM estimator, we set the weight matrix as $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ for the efficient MD estimator. Then it follows that $-\mathbf{H} = \boldsymbol{\Sigma} = \mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G} \equiv \mathbf{V}$ and $\mathbf{V}_{\boldsymbol{\Pi}_s} = \boldsymbol{\Pi}_s'(\boldsymbol{\Pi}_s\mathbf{V}\boldsymbol{\Pi}_s')^{-1}\boldsymbol{\Pi}_s$. Therefore, the FIC for the efficient MD estimator has the same form as (S3.10).

S4. Additional numerical results

Figure 1 presents the model weights of W-opt, PIA-1, and PIA-2 placed on each submodel. For W-opt, the model weights are calculated based on (3.18) for each d. For PIA-1 and PIA-2, we calculate $\mathrm{E}(w^{\infty})$ based on Corollary 2 by simulation averaging across 10,000 random samples. The numerical results show that W-opt assigns more weights to the narrow/full model for smaller/larger $|d|$, which is consistent with the relative performance between Narrow and Full displayed in Figure 1. Similar to W-opt, both PIA-1 and PIA-2 put more weights on the narrow/full model when $|d|$ is small/large. However, compared to W-opt, both PIA-1 and PIA-2 tend to assign more weights to the middle model for a fixed value of d, which is not optimal. Therefore, PIA-1 and PIA-2 have larger AMSEs than W-opt as shown in Figure 1.

Figures 2 and 3 present the relative MSEs of different estimates in the heteroskedastic setup for $n = 100$ and 250, respectively. Similar to the results in the homoskedastic setup, the relative performance of these estimators depends strongly on c, p, and S. When the number of must-have parameters p increases or the number of submodels S decreases, the relative MSEs of these estimators are getting close to each other. Overall, the ranking of different estimators in the heteroskedastic setup is quite

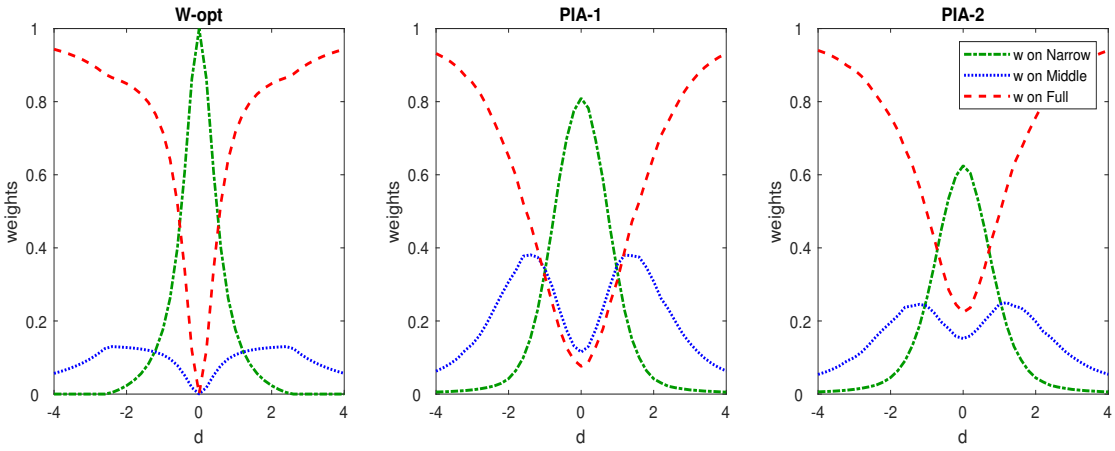

Figure 1: Model weights placed on each submodels

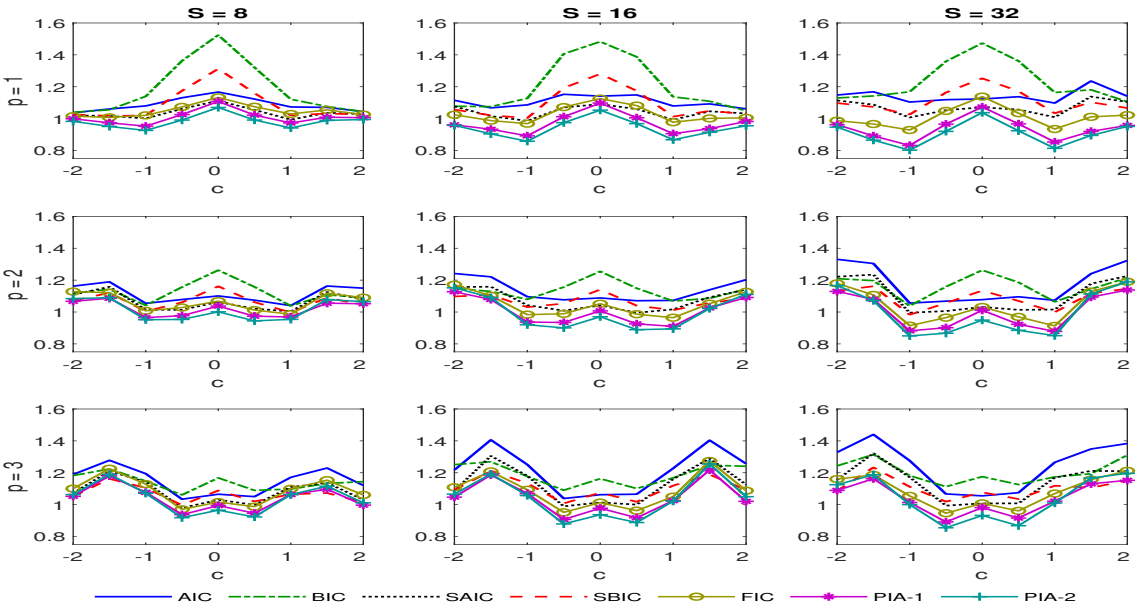

Figure 2: Relative MSE, heteroskedastic errors, $n = 100$.

similar to that in the homoskedastic setup, and PIA-2 still achieves a lower MSE than other estimators in most cases.

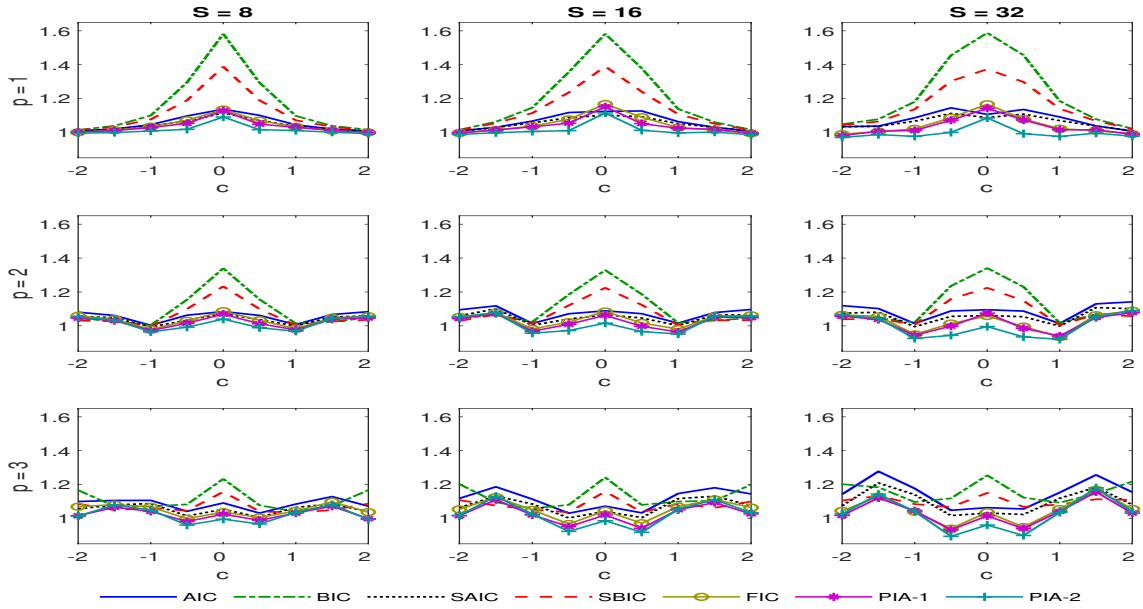

Figure 3: Relative MSE, heteroskedastic errors, $n = 250$.

S5. Post-averaging inference

Let $w(s|\widehat{\boldsymbol{\delta}})$ denote a data-dependent weight function for the sth submodel. Consider an averaging estimator of the focus parameter μ_0 as

$$\widehat{\mu} = \sum_{s=1}^{S} w(s|\widehat{\boldsymbol{\delta}})\widehat{\mu}_s, \tag{S5.1}$$

where the weight $w(s|\widehat{\boldsymbol{\delta}})$ takes the value in the interval $[0, 1]$ and the sum of weights equals 1. Suppose that $w(s|\widehat{\boldsymbol{\delta}}) \xrightarrow{d} w(s|\boldsymbol{\Delta})$, where $\boldsymbol{\Delta} = \boldsymbol{\delta}_0 + \boldsymbol{\Pi}_0'\mathbf{Z}$. The following theorem presents a general distribution theorem for the averaging estimator with data-dependent weights.

Theorem A1. Suppose that Assumptions 1–3 hold. Assume $w(s|\widehat{\boldsymbol{\delta}}) \xrightarrow{d}$

$w(s|\boldsymbol{\Delta})$ with at most a countable number of discontinuities. As $n \to \infty$, we have

$$\sqrt{n}(\widehat{\mu} - \mu_0) \xrightarrow{d} \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{Z} + \mathbf{D}_{\boldsymbol{\theta}}' \left(\sum_{s=1}^{S} w(s|\boldsymbol{\Delta}) \mathbf{B}_s \right) \boldsymbol{\Pi}_0 \boldsymbol{\Delta},$$

where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1})$ and $\mathbf{B}_s = \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q}$.

Unlike Theorem 2, Theorem A1 shows that the averaging estimator with data-dependent weights has a nonstandard asymptotic distribution since the estimated weights are asymptotically random. This nonstandard asymptotic distribution can be expressed in terms of a nonlinear function of the normal random vector $\mathbf{Z}$.

We follow Hjort and Claeskens (2003), Claeskens and Carroll (2007), and Zhang and Liang (2011) to construct a valid confidence interval as follows. Let $\widehat{\kappa}^2 = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}'\widehat{\mathbf{H}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{H}}^{-1}\widehat{\mathbf{D}}_{\boldsymbol{\theta}}$, which is a consistent estimator of $\mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}\mathbf{D}_{\boldsymbol{\theta}}$. Recall that $\widehat{\boldsymbol{\delta}} \xrightarrow{d} \boldsymbol{\Delta} \sim N(\boldsymbol{\delta}_0, \boldsymbol{\Pi}_0'\mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1}\boldsymbol{\Pi}_0)$. From Theorem A1, it is easy to see that

$$\left[\sqrt{n}(\widehat{\mu} - \mu_0) - \widehat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\sum_{s=1}^{S} w(s|\widehat{\boldsymbol{\delta}}) \widehat{\mathbf{B}}_s \right) \boldsymbol{\Pi}_0 \widehat{\boldsymbol{\delta}} \right] / \widehat{\kappa} \xrightarrow{d} N(0,1). \tag{S5.2}$$

Let $b(\widehat{\boldsymbol{\delta}}) = \widehat{\mathbf{D}}_{\boldsymbol{\theta}}' \left(\sum_{s=1}^{S} w(s|\widehat{\boldsymbol{\delta}}) \widehat{\mathbf{B}}_s \right) \boldsymbol{\Pi}_0 \widehat{\boldsymbol{\delta}}$. Then, we can construct the confidence interval for μ_0 as

$$\mathrm{CI}_n = \left[\widehat{\mu} - \frac{b(\widehat{\boldsymbol{\delta}})}{\sqrt{n}} - z_{1-\alpha/2}\frac{\widehat{\kappa}}{\sqrt{n}},\ \widehat{\mu} - \frac{b(\widehat{\boldsymbol{\delta}})}{\sqrt{n}} + z_{1-\alpha/2}\frac{\widehat{\kappa}}{\sqrt{n}} \right], \tag{S5.3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. From (S5.2), we have $Pr(\mu_0 \in \mathrm{CI}_n) \to 2\Phi(z_{1-\alpha/2}) - 1$, where $\Phi(\cdot)$ is a standard normal distribution function, which means the proposed confidence interval (S5.3) has asymptotically the correct coverage probability.

Proof of Theorem A1: Since all of Λ_s can be expressed in terms of the same normal vector $\mathbf{Z}$ in Corollary 1, there is joint convergence in distribution of all $\sqrt{n}(\widehat{\mu}_s - \mu_0)$ to Λ_s for $s = 1, \ldots, S$, Also, $w(s|\widehat{\boldsymbol{\delta}}) \xrightarrow{d} w(s|\boldsymbol{\Delta})$, where $w(s|\boldsymbol{\Delta})$ is a function of the random vector $\mathbf{Z}$. Recall that $\mathbf{B}_s = \mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q}$. Therefore,

$$
\begin{aligned}
&\sqrt{n}(\widehat{\mu} - \mu_0) \\
&= \sum_{s=1}^{S} w(s|\widehat{\boldsymbol{\delta}})\sqrt{n}(\widehat{\mu}_s - \mu_0) \\
&\xrightarrow{d} \sum_{s=1}^{S} w(s|\boldsymbol{\Delta}) \left(\mathbf{D}_{\boldsymbol{\theta}}'(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H} - \mathbf{I}_{p+q})\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z}\right) \\
&= \mathbf{D}_{\boldsymbol{\theta}}' \sum_{s=1}^{S} w(s|\boldsymbol{\Delta})(\mathbf{B}_s\boldsymbol{\Pi}_0\boldsymbol{\delta}_0 + \mathbf{B}_s\boldsymbol{\Pi}_0\boldsymbol{\Pi}_0'\mathbf{Z}) + \mathbf{D}_{\boldsymbol{\theta}}' \sum_{s=1}^{S} w(s|\boldsymbol{\Delta})(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}\mathbf{Z} - \mathbf{B}_s\boldsymbol{\Pi}_0\boldsymbol{\Pi}_0'\mathbf{Z}) \\
&= \mathbf{D}_{\boldsymbol{\theta}}' \sum_{s=1}^{S} w(s|\boldsymbol{\Delta})\mathbf{B}_s\boldsymbol{\Pi}_0(\boldsymbol{\delta}_0 + \boldsymbol{\Pi}_0'\mathbf{Z}) + \mathbf{D}_{\boldsymbol{\theta}}' \sum_{s=1}^{S} w(s|\boldsymbol{\Delta})(\mathbf{H}_{\boldsymbol{\Pi}_s}\mathbf{H}(\mathbf{I}_{p+q} - \boldsymbol{\Pi}_0\boldsymbol{\Pi}_0') + \boldsymbol{\Pi}_0\boldsymbol{\Pi}_0')\mathbf{Z} \\
&= \mathbf{D}_{\boldsymbol{\theta}}' \left(\sum_{s=1}^{S} w(s|\boldsymbol{\Delta})\mathbf{B}_s \right) \boldsymbol{\Pi}_0\boldsymbol{\Delta} + \mathbf{D}_{\boldsymbol{\theta}}'\mathbf{Z},
\end{aligned}
$$

where the last equality holds by the facts that $\boldsymbol{\Delta} = \boldsymbol{\delta}_0 + \boldsymbol{\Pi}_0'\mathbf{Z}$ and

$$\mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}(\mathbf{I}_{p+q} - \mathbf{\Pi}_0\mathbf{\Pi}_0') = \mathbf{H}_{\mathbf{\Pi}_s}\mathbf{H}\begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times q} \end{bmatrix}$$

$$= \mathbf{\Pi}_s'(\mathbf{\Pi}_s\mathbf{H}\mathbf{\Pi}_s')^{-1}\mathbf{\Pi}_s\mathbf{H}\mathbf{\Pi}_s'\begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q_s\times p} & \mathbf{0}_{q_s\times q} \end{bmatrix} = \mathbf{I}_{p+q} - \mathbf{\Pi}_0\mathbf{\Pi}_0'.$$

This completes the proof. ■

References

Chang, M. and F. J. DiTraglia (2018). A generalized focused information criterion for GMM. *Journal of Applied Econometrics 33*(3), 378–397.

Claeskens, G. and R. J. Carroll (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika 94*, 249–265.

Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association 98*(464), 900–916.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Hansen, B. E. (2022). *Econometrics*. Princeton University Press.

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

Hjort, N. L. and G. Claeskens (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association 101*(476), 1449–1464.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In

R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier.

Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer Verlag.

Wang, H. and C. Leng (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association 102*(479), 1039–1048.

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