Statistica Sinica **34** (2024), 771-792
doi:https://doi.org/10.5705/ss.202021.0266

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

Xinyu Zhang and Chu-An Liu*

Chinese Academy of Sciences and Academia Sinica

Abstract: Traditional model selection criteria select a single model based on its global fit of the model. In contrast, a focused information criterion is tailored to a parameter of interest, based on which, it selects a model. We propose 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. The results from Monte Carlo simulations and an analysis of real data show that the proposed selection and averaging methods perform comparably with other methods.

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

1. Introduction

There is a long history of model selection methods in the econometric and statistical literature. Traditional model selection criteria, such as the Akaike information criterion and Bayesian information criterion, choose a 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. For example, Hansen (2005) shows that finite-sample optimal model selection may be sensitive to the choice of parameter of interest. Claeskens, Croux and Van Kerckhoven (2006) provide specific examples in biostatistics in which no single model is good for every patient subgroup. Instead of choosing a single model to explain all aspects of the data, a focused information criterion (FIC; Claeskens and Hjort (2003)) selects a model based on a parameter of interest, each of which can have its own model.

Since the seminal work of Claeskens and Hjort (2003), the FIC has been investigated for various 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 error 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

*Corresponding author.

mixed-effects model (Chen, Zou and Zhang (2013)), generalized empirical likelihood estimation (Sueishi (2013)), the graphical model (Pircalabelu, Claeskens and Waldorp (2015)), the propensity score-weighted estimation of 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)), and vector autoregressive models (Lohmeyer et al. (2019)), among others. It is well known that many of these estimators share a common structure, which is useful when deriving the FIC in different model setups. Therefore, it would be interesting to know whether it is feasible to develop an FIC for various models in a unified theoretical framework, instead of in a case-by-case manner.

In this paper, we propose an 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 a parameter of interest 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 an 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 parameter estimate. Thus, it selects the model that minimizes the estimated AMSE. We apply our results to several examples, and provide the FIC in each case, including the nonlinear least squares (NLS) estimator, maximum likelihood estimator, generalized method of moments estimator, and 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 using fixed weights for the parameter of interest, 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 for the parameter of interest.

We investigate the asymptotic and finite-sample properties of the proposed FIC and plug-in averaging method. We show that both the FIC and the estimated weights are asymptotically random under the local asymptotic framework. 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 the asymptotic behavior of the FIC and the plug-in averaging method. Using simulations, we compare the finite-sample performance of the proposed method with that of existing model selection and model averaging methods. In a real-data analysis, we apply the proposed method to investigate the relationship between income and education. Our simulation studies and empirical results both show that the proposed method performs well and achieves lower mean squared errors (MSEs), in general, than those of 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 the 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 Material.

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 the data and the sample size n. We consider a general class of estimators, referred to as extremum estimators by Newey and McFadden (1994), that maximizes this objective function. Note that $\widehat{Q}_n(\boldsymbol{\theta})$ can be a negative log-likelihood function, a least squares function, a minimum-distance criterion function, and so on. For example, if we set $\widehat{Q}_n(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n m_{\boldsymbol{\theta}}(\mathbf{x}_i)$, where $m_{\boldsymbol{\theta}}(\cdot)$ is a real-value function of $\mathbf{x}_i$, then the extremum estimator is an M-estimator that includes the maximum likelihood estimator and NLS 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 $g_n(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^n g(\mathbf{z}_i, \boldsymbol{\theta})$ is a sample average of the moment functions, then the extremum estimator is the generalized method of moments estimator.

Our goal is to select a model based on a parameter of interest in a general setting that allows for parameter uncertainty. In our framework, the candidate models can be nested or non-nested, and we are uncertain about which model parameters should be included in each candidate 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 $\boldsymbol{\gamma}$. Because the true value of $\boldsymbol{\gamma}$ may be zero, we 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 $\boldsymbol{\gamma}$, whereas for the narrow model, we include only $\boldsymbol{\beta}$, and set all $\boldsymbol{\gamma}$ to zero. We can also set some $\boldsymbol{\gamma}$ to zero, 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 $\boldsymbol{\gamma}_{s^c}$ be 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

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \widehat{Q}_n(\boldsymbol{\theta}). \tag{2.1}$$

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

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

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

We now state the regularity conditions required for the asymptotic results, where all limiting processes 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

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

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)* $\sqrt{n}(\partial\widehat{Q}_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$. *(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 the 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 discussion 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*

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

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 is similar to Condition C4 in Hjort and Claeskens (2003) and Condition A4 in Claeskens and Carroll (2007). However, we exclude the quantile regression model from our framework, owing to the failure of differentiability. For focused information criteria 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 technique ensures that the AMSE 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 a model selection estimator, for example, as in Hjort and Claeskens (2003), Leeb and Pötscher (2005), and Claeskens and Hjort (2008). This assumption implies that the submodels become more similar as the sample size increases, and yields the same stochastic order of squared biases and variances. Hence, the optimal model achieves the best trade-off between the bias and variance in this context. As an alternative, other works assume that the parameters decay at 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 setup of asymptotics with fixed parameters, 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 in a local neighborhood of zero. Let $\mathbf{I}$ denote the 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*

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

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 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) simplifies 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, respectively, of the submodel estimator. Our theorem demonstrates that the trade-off between the squared biases and variances remains in the asymptotic theory, a feature that is essential to the proposed FIC and plug-in averaging method.

Remark 2. The proof of Theorem 1 is not a trivial extension of existing results. Note 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 to those in Fan and Li (2001) and Wang and Leng (2007) to show that $\widehat{\boldsymbol{\theta}}_s - \widehat{\boldsymbol{\theta}} = O_p(n^{-1/2})$. We then show that $\widehat{\boldsymbol{\theta}}_s$ is approximately 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. Therefore, the asymptotic distribution of the submodel estimator is a linear function of the normal random vector $\mathbf{Z}$.

3. FIC and Plug-In Averaging Method

In this section, we propose an FIC for extremum estimators. As an illustration, we apply the general results to the NLS estimator. We also provide additional examples to illustrate the general results in the supplementary materials, including the maximum likelihood estimator, generalized method of moments estimator, and 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, rather than assessing the overall properties of the model. Unlike traditional model selection approaches that assess the global fit of a model, we evaluate the model based on the parameter of interest. Let $\mu = \mu(\boldsymbol{\theta}) = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$ be the parameter of interest,

which is a smooth real-valued function. Note that if μ depends only on $\boldsymbol{\gamma}$, and the estimator sets $\boldsymbol{\gamma} = \mathbf{0}$, 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 parameter of interest 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 the partial derivatives evaluated at the null points $\boldsymbol{\theta}_0^*$, that is,

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

We select the 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*

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

From Corollary 1, a direct calculation yields

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

Because $\mathbf{D}_{\boldsymbol{\theta}}$ depends on the parameter of interest μ, we can use (3.2) to select a proper submodel that depends on this parameter. 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 their sample analogs. The proposed FIC of the sth submodel is defined as

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

which is an asymptotically unbiased estimator of the MSE $\mathrm{E}(\Lambda_s^2)$, in the sense that the mean of the asymptotic distribution of FIC_s is equal to the MSE $\mathrm{E}(\Lambda_s^2)$. Here, $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is defined in (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 $\widehat{\boldsymbol{\theta}}^* = (\widehat{\boldsymbol{\beta}}', \mathbf{0}')'$. Because $\widehat{\boldsymbol{\theta}}$ 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 estimate $\mathbf{H}$ consistently by using the sample analog $\widehat{\mathbf{H}}$ under Assumption 1. Similarly, the covariance

matrix $\boldsymbol{\Sigma}$ can be estimated consistently using 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 local asymptotic framework means a consistent estimator for $\boldsymbol{\delta}_0$ is not available. However, we can construct an asymptotically unbiased 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: NLS estimator

Suppose the data $(y_i, \mathbf{x}_i')'$ are independent and identically distributed (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 the 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 $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}_0) = -\mathrm{E}((\partial h(\mathbf{x}_i, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta})(\partial h(\mathbf{x}_i, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}'))$. 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

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

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 $\boldsymbol{\Sigma}$, respectively, and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is an asymptotically unbiased estimator of $\boldsymbol{\delta}_0\boldsymbol{\delta}_0'$.

When the error term e_i is homoskedastic, that is, $\mathrm{E}(e_i^2|\mathbf{x}_i) = \sigma^2$, we have $\boldsymbol{\Sigma} = -\sigma^2\mathbf{H}$, and the covariance matrix $\mathbf{V}_{\boldsymbol{\Pi}_s}$ simplifies to $-\sigma^2\mathbf{H}_{\boldsymbol{\Pi}_s}$. 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 for the parameter of interest. 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 FIC_s, and gives the other models zero weights. Thus, the weight function of the FIC is $\widehat{w}_s = \mathbf{1}\{\mathrm{FIC}_s = \min(\mathrm{FIC}_1, \mathrm{FIC}_2, \ldots, \mathrm{FIC}_S)\}$, where $\mathbf{1}\{\cdot\}$ is an indicator function that takes a value of either zero or one.

We now consider a general weight function, instead of a zero-one weight function. Rather than comparing the AMSE of each submodel, we first derive the AMSE of the averaging estimator with a 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

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

Theorem 2 shows the asymptotic normality of the averaging estimator with fixed weights, and 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 equal to

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

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

$$\mathbf{w}^{\text{opt}} = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\Psi}\mathbf{w}, \tag{3.18}$$

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 $\widehat{\mu}(\mathbf{w})$.

However, the optimal weight vector is infeasible, because $\boldsymbol{\Psi}$ is unknown. Here, we can obtain a feasible version of $\mathbf{w}^{\text{opt}}$ by replacing the unknown parameters in $\boldsymbol{\Psi}$ with their sample analogs. As discussed in Section 3.1, the unknown parameters $\mathbf{D}_{\boldsymbol{\theta}}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$ can be estimated consistently by their sample analogs. Note that a consistent estimator for $\boldsymbol{\delta}_0$ is not available because of the local-to-zero assumption. Therefore, following Wan, Zhang and Wang (2014) and Liu (2015), we propose the following plug-in estimator of $A(\mathbf{w})$:

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

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

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

and $\widehat{\boldsymbol{\delta}\boldsymbol{\delta}'}$ is defined in (3.5). Note 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

$$\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_S)' = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}}\, \mathbf{w}' \widehat{\boldsymbol{\Psi}} \mathbf{w}. \tag{3.21}$$

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 determined numerically using 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) may have better asymptotic and finite-sample properties than those of 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) does 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), and 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*

$$\widehat{\mathbf{w}} \xrightarrow{d} \mathbf{w}^{\infty} = \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}}\, \mathbf{w}' \mathbf{\Psi}^{\infty} \mathbf{w} \tag{3.25}$$

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) do 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 $\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'$ 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) and 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 further details.

Remark 3. 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, $\mathbf{w}' \mathbf{\Psi}^{\infty} \mathbf{w}$ has a unique minimum; see Charkhi, Claeskens and Hansen (2016) for a discussion on the uniqueness of the weights. For the estimator defined in (3.23), the estimated weights are still random in the limit, because we can show that

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

Compared with (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. Next, we use Monte Carlo experiments to compare the finite-sample performance of the proposed method with that of existing model selection and model averaging methods. In the last subsection, we apply the proposed method to a real-data analysis.

4.1. AMSE comparison

We evaluate the asymptotic performance of the different estimates of the parameter of interest μ based on the numerical calculation of the AMSE. We consider a simple three-nested-model framework based on 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 a homoskedastic error, and set $\sigma^2 = 1$ and $\boldsymbol{\Sigma} = -\sigma^2\mathbf{H}$, where the diagonal elements of $\mathbf{H}$ are -1, and the off-diagonal elements are -0.5. The parameter of interest is $\mu = \theta_1$, and $\mathbf{D}_{\boldsymbol{\theta}} = (1, 0, 0)'$ in this setting. We compare the AMSE of the following estimators: (1) a narrow model estimator (labeled Narrow); (2) a middle model estimator (labeled Middle); (3) the full model estimator (labeled Full); (4) an averaging estimator with the optimal weights $\mathbf{w}^{\text{opt}}$ defined in (3.6) (labeled W-opt); (5) the FIC model selection estimator (labeled FIC); (6) the plug-in averaging method based on (3.20) (labeled PIA-1); and (7) the 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, and 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 using (3.18), and then calculate the AMSE by plugging the values 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 ease of comparison. If the relative AMSE exceeds one, then the specified estimator has a larger AMSE than that of the averaging estimator with the optimal weights.

Figure 1 presents the relative AMSEs for the various estimators. We first compare the AMSEs between the submodel estimators and W-opt. As expected, the narrow model achieves a lower relative AMSE than the other two submodels do for smaller $|d|$, whereas the full model achieves a smaller relative AMSE than

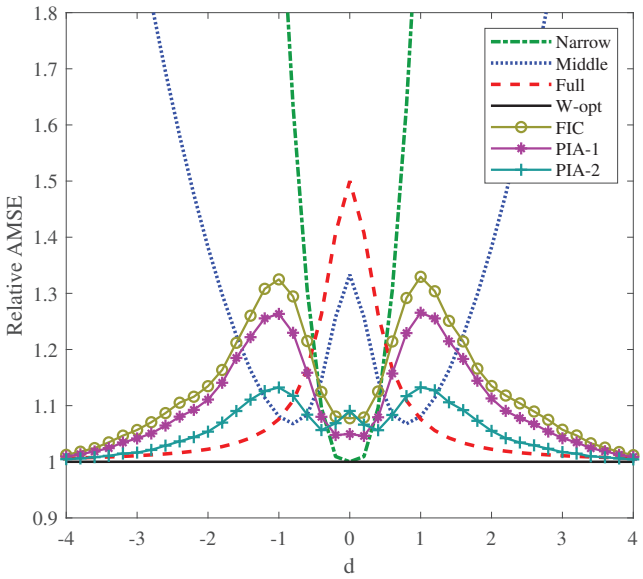

Figure 1. Relative AMSE.

the other two submodels do for larger $|d|$. Therefore, the best submodel with the lowest AMSE varies with d. Compared with 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 that of PIA-1, and PIA-1 has a smaller relative AMSE than that of the FIC. Note that the AMSE of PIA-2 is slightly larger than that of W-opt, showing the effect of the estimated local parameter on the asymptotic behavior of the plug-in averaging method. Similarly, for a fixed value of d, the AMSE of the FIC is larger than that of the best submodel, owing 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 method using Monte Carlo experiments. We consider the following nonlinear regression model:

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

where $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})' \sim i.i.d.$ Uniform$(-1, 1)$ and $\mathbf{z}_i = (z_{1i}, \ldots, z_{qi})' \sim i.i.d.$ 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 let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ be 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 we 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 the FIC, PIA-1, and PIA-2, we also consider the following estimators: (1) the Akaike information criterion model selection estimator (labeled AIC); (2) the Bayesian information criterion model selection estimator (labeled BIC); (3) a smoothed AIC model selection estimator (labeled SAIC); and (4) a smoothed BIC model selection estimator (labeled SBIC). Let $\widehat{\sigma}_s^2 = (1/n)\sum_{i=1}^n \widehat{e}_{si}^2$, where $\widehat{e}_{si}$ is the NLS residual from the model s. The AIC of the sth model is $\mathrm{AIC}_s = n\log(\widehat{\sigma}_s^2) + 2(p+q_s)$, where $p+q_s$ is the number of parameters in the model s, and the BIC of the sth model is $\mathrm{BIC}_s = n\log(\widehat{\sigma}_s^2) + \log(n)(p+q_s)$. For the AIC and BIC, we select the model with the lowest value. The SAIC estimator is proposed by Buckland, Burnham and Augustin (1997), and uses the exponential AIC as the model weight. The SAIC weight is proportional to the likelihood of the model, and is defined as $\widehat{w}_s = \exp(-(1/2)\mathrm{AIC}_s)/\sum_{r=1}^S \exp(-(1/2)\mathrm{AIC}_r)$. The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors, and the SBIC weight is $\widehat{w}_s = \exp(-(1/2)\mathrm{BIC}_s)/\sum_{r=1}^S \exp(-(1/2)\mathrm{BIC}_r)$.

Our parameter of interest is $\mu = \beta_p$, which is the last element of the must-have parameters. To evaluate the finite-sample behavior of each estimator, we compare these estimators based on the MSE of $\widehat{\mu}$. The MSE is calculated as the average of $(\widehat{\mu} - \mu)^2$ obtained from each method over 5,000 replications. For ease of comparison, we divide the MSE of each method by that of the best-fitting submodel, and report the relative MSE in each case. The best-fitting submodel has the lowest MSE among all submodels. Therefore, a 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.

Figures 2 and 3 present the relative MSEs of the various 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 the AIC, BIC, SAIC, and SBIC. The simulation results show that the BIC has a larger MSE than that of the AIC for smaller $|c|$ in all cases, and that the AIC has a larger MSE than that of the BIC for larger $|c|$ when $p = 2$ and 3. The SAIC and SBIC have lower MSEs than those of the AIC and BIC, respectively, and the pattern of relative performance between the SAIC and the SBIC is quite similar to that between the AIC and the BIC. We next compare the finite-sample performance of the FIC, PIA-1, and

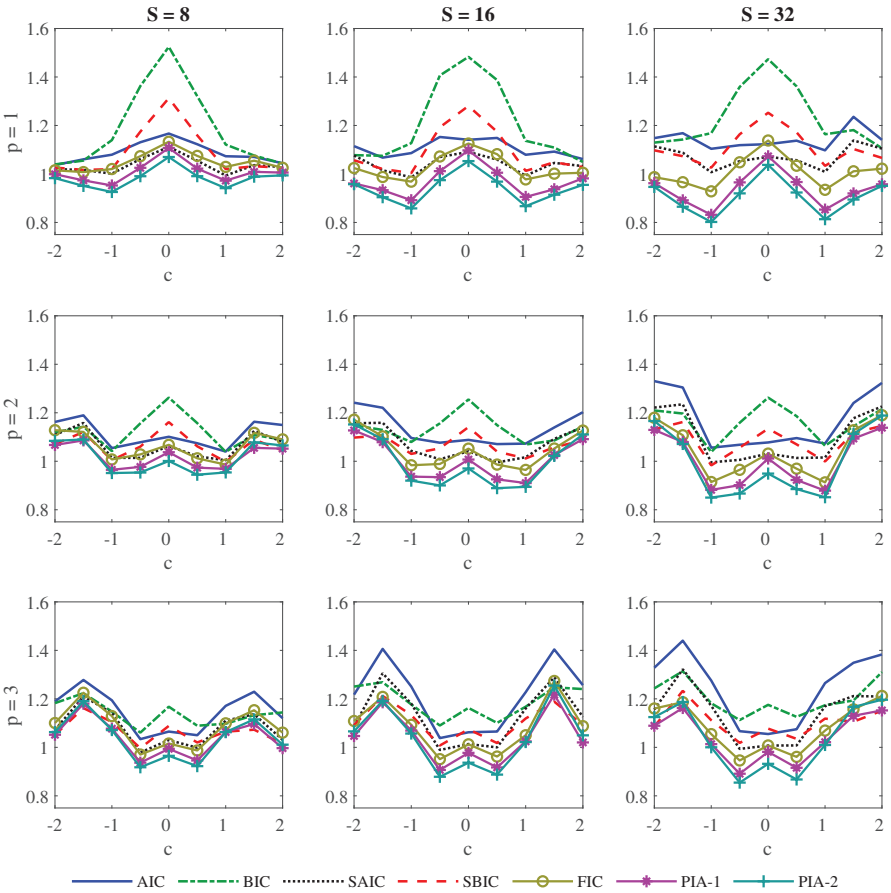

Figure 2. Relative MSE, homoskedastic errors, $n = 100$.

PIA-2. The results show that all three perform quite well, and have lower MSEs than those of the AIC, BIC, SAIC, and SBIC in most cases. The PIA-2 performs slightly better than the PIA-1, and the PIA-1 performs slightly better than the FIC. The relative performance of the 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 method to investigate the relationship between income and education. We use the German Socioeconomic Panel data set of Riphahn, Wambach and Million (2003) 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 from 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 income. After deleting two observations

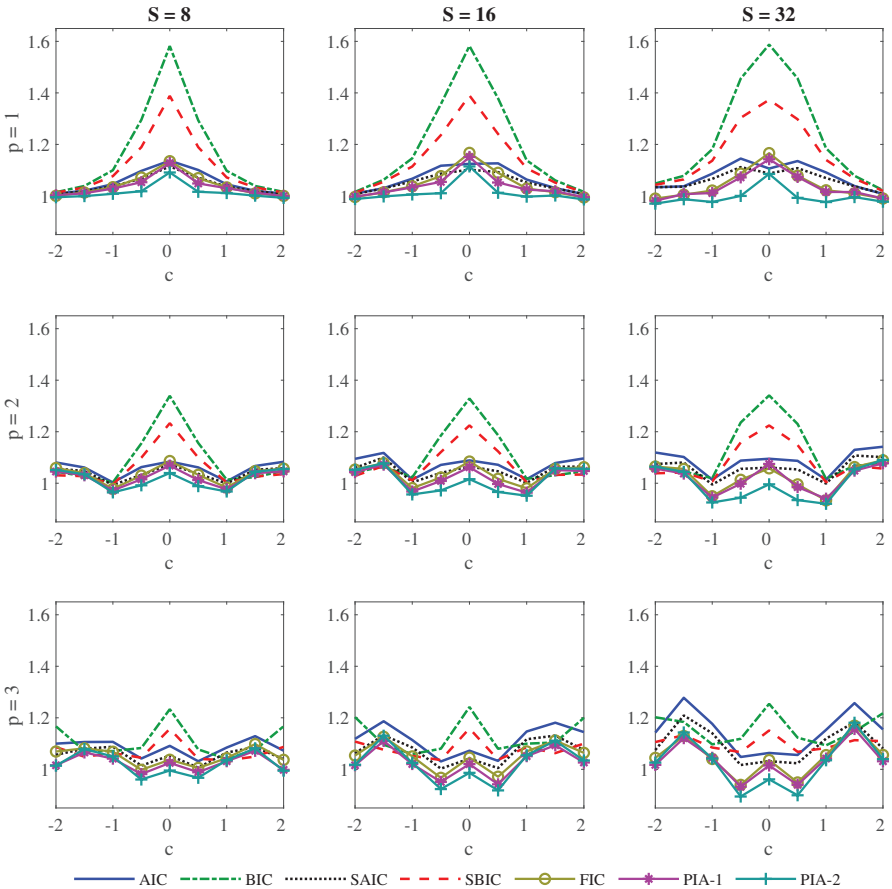

Figure 3. Relative MSE, homoskedastic errors, $n = 250$.

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 = female, 0 = male), and the quadratic and interaction terms of the 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 the variables as potentially relevant regressors. We consider all possible subsets of potentially relevant regressors, yielding 32 non-nested models. Our parameter of interest 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. 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.

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

Following Rolling, Yang and Velez (2019), we perform a guided simulation experiment to evaluate the different methods under simulation scenarios that are consistent with the data. The simulation scenario is based on the submodel selected by the 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 includes only 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 $\epsilon_i \sim i.i.d.$ 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 parameter of interest μ, that is, the coefficient of Education. Note that the true value of μ is known for each choice of τ. From Table 1, the true

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

values of μ are 0.1249, 0.1217, and 0.1279 for the AIC, BIC, and FIC scenarios, respectively.

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

5. Conclusion

We have investigated the limiting distribution of extremum estimators in a local asymptotic framework, and propose an FIC and a plug-in averaging method for extremum estimators. We have also investigated the asymptotic and finite-sample properties of the proposed method. We find that the FIC model selection estimator and the averaging estimator with data-driven weights have nonstandard limiting distributions, owing 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 values than those of existing methods.

Supplementary Material

The online supplementary material includes proofs, additional examples, and numerical results, as well as details on how to construct a valid confidence interval for the post-averaging estimator.

Acknowledgments

We thank the editor, associate editor, and 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., Claeskens, G. and Dette, H. (2014). Focused model selection in quantile regression. *Statistica Sinica* **24**, 601–624.

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

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

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

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

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

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

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

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

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

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

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

Greene, W. H. (2012). *Econometric Analysis*. 7th Edition. Pearson.

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

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

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

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

1464.

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

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

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

Leeb, H. and Pötscher, B. (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., Palm, F., Reuvers, H. and Urbain, J.-P. (2019). Focused information criterion for locally misspecified vector autoregressive models. *Econometric Reviews* **38**, 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**, 46–75.

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

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

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

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

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature* **58**, 644–719.

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

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

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

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

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

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

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

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

Xinyu Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100080, China.
E-mail: xinyu@amss.ac.cn
Chu-An Liu
Institute of Economics, Academia Sinica, Taipei City 115, Taiwan.
E-mail: caliu@econ.sinica.edu.tw

(Received July 2021; accepted August 2022)