<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2021-0266</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Authors</strong></td>
</tr>
<tr>
<td><strong>E-mails</strong></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
A UNIFIED APPROACH TO FOCUSED INFORMATION CRITERION AND PLUG-IN AVERAGING METHOD

Xinyu Zhang\textsuperscript{1,2} and Chu-An Liu\textsuperscript{3,*}

\textsuperscript{1}Academy of Mathematics and Systems Science, Chinese Academy of Sciences
\textsuperscript{2}Beijing Academy of Artificial Intelligence
\textsuperscript{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 \cite{Claeskens2003} 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.

\textsuperscript{*}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 \cite{Wang2012}, the Tobin model with a nonzero threshold \cite{Zhang2012}, the partially linear single-index model \cite{Yu2013}, the linear mixed-effects model \cite{Chen2013}, generalized empirical likelihood estimation \cite{Sueishi2013}, the graphical model \cite{Pircalabelu2015}, propensity score weighted estimation of the treatment effects \cite{Lu2015, Kitagawa2016}, the choice between parametric and nonparametric models \cite{Jullum2017}, generalized method of moments estimation \cite{DiTraglia2016, Chang2018}, vector autoregressive models \cite{Lohmeyer2019}, 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 \cite{Newey1994}, 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 pro-
posed 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 distribu-
tions. 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 in-
troduces 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 $\theta = (\beta', \gamma')' \in \Theta \subset \mathbb{R}^{p+q}$ denote a $p+q$ vector of unknown parameters, where $\Theta$ is the set of possible parameter values. Suppose we have a sample objective function $Q_n(\theta)$ 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 $Q_n(\theta)$ could be a negative log-likelihood function, a least squares function, a minimum-distance criterion function, and so on. For example, if we set $Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i)$, where $m_\theta(\cdot)$ is a real-value function of $x_i$, 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 $Q_n(\theta) = -g_n(\theta)' W_n g_n(\theta)$, where $W_n$ is a positive semi-definite weight matrix and $g_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta)$ 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 $\beta$ is a $p \times 1$ vector of “must-have” parameters that must be included in the model based on theoretical grounds, and $\gamma$ 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 $s$th submodel includes all $\beta$ but some or none of the components $\gamma$. Since the true value of $\gamma$ could be zeros, we could restrict some elements of $\gamma$ 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 $\gamma$, then we have $S = 2^q$ submodels.

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

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{Q}_n(\theta). \quad (2.1)$$
For the $s$th submodel, the unknown parameters are $\eta_s = (\beta', \gamma_s')'$. Let $\Pi_s$ be a $(p + q_s) \times (p + q)$ projection matrix such that $\Pi_s \theta = \eta_s$, where $q_s$ is the dimension of $\gamma_s$. Similarly, let $\Pi_{sc}$ be a projection matrix such that $\Pi_{sc} \theta = \gamma_{sc}$. Hence, we can write $\widehat{Q}_n(\theta)$ as $\widehat{Q}_n(\beta, \gamma_s, \gamma_{sc})$ and the extremum estimator for the $s$th submodel is

$$\widehat{\theta}_s = \Pi_s' \widehat{\eta}_s = \arg \max_{\theta \in \Theta} \widehat{Q}_n(\beta, \gamma_s, 0),$$

(2.2)

where $0$ is a zero vector. Note that $\widehat{\theta}_s$ is a $(p + q) \times 1$ vector with the values of $\gamma_{sc}$ 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 $\widehat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, and $Q_0(\theta)$ is uniquely maximized at $\theta_0 = (\beta'_0, \gamma'_0)'$. Define $\theta'_0 = (\beta'_0, 0')'$ as the null points. Let

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

be the Hessian matrix of second derivatives and the expected Hessian matrix, respectively. Let $\overset{p}{\to}$ and $\overset{d}{\to}$ denote convergence in probability and convergence in distribution, respectively. Let $\| \cdot \|$ denote the Euclidean norm.
Assumption 1. (i) $\hat{\theta} - \theta_0 \overset{p}{\to} 0$. (ii) $\theta_0$ is in the interior of $\Theta$. (iii) $Q_n(\theta)$ is twice continuously differentiable in a neighborhood $\Theta_0 \subset \Theta$ of $\theta_0$. (iv) $\sqrt{n} \frac{\partial}{\partial \theta} Q_n(\theta_0) \overset{d}{\to} N(0, \Sigma)$. (v) There is $H(\theta)$ that is continuous at $\theta_0$ for every $n$, and $\sup_{\theta \in \Theta} \| H_n(\theta) - H(\theta) \| \overset{p}{\to} 0$. (vi) $H(\theta_0)$ 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 $\hat{\theta}$, and this condition holds under appropriate primitive assumptions; see the discussions in Section 2 of Newey and McFadden (1994). Let $H = H(\theta_0)$. Under Assumption 1, Theorem 3.1 of Newey and McFadden (1994) demonstrates the asymptotic normality of $\hat{\theta}$:

$$Z_n \equiv \sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\to} Z \sim N(0, H^{-1} \Sigma H^{-1}), \quad (2.3)$$

where $Z$ is a normal random vector and $\Sigma$ is a positive definite matrix.

Assumption 2. $\hat{Q}_n(\theta)$ is three times differentiable in a neighborhood $\Theta_0^* \subset \Theta$ of $\theta_0^*$, and the third partial derivative of $\hat{Q}_n(\theta)$ satisfies

$$\sup_{\theta_0^* \in \Theta_0^*} \frac{\partial^3 \hat{Q}_n(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}|_{\theta = \theta_0^*} = o_p(n^{1/2}).$$

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.** $\gamma_0 \equiv \gamma_{0,n} = \delta_0 / \sqrt{n}$ where $\delta_0$ is an unknown constant vector.

Assumption 3 specifies that $\gamma_0$ is in a local $n^{-1/2}$ neighborhood of zero, and thus $\theta_0 = (\beta_0', \delta_0' / \sqrt{n})'$. 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 $\gamma$ being a local neighborhood of zero. Let $I$ 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

$$\sqrt{n}(\hat{\theta}_s - \theta_0^*) \overset{d}{\to} \Pi_s^\prime H_{\Pi_s} H_{\Pi_0} \delta_0 + H_{\Pi_s} H Z \sim N(H_{\Pi_s} H_{\Pi_0} \delta_0, H_{\Pi_s} \Sigma H_{\Pi_s}), \quad (2.4)$$

where $H_{\Pi_s} = \Pi_s^\prime (H_{\Pi_s} H_{\Pi_s}^\prime)^{-1} \Pi_s$ and $\Pi_0 = (0_{q \times p}, I_q)'$.

**Remark 1.** Theorem 1 extends the asymptotic theory of extremum estimators for drifting sequences of parameters, and it implies that the submodel estimator $\hat{\theta}_s$ is root-$n$ consistent. When we set $\Pi_s = I_{p+q}$ for the full model, we have $\hat{\theta}_s = \hat{\theta}$. 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, $H_{\Pi_s} H_{\Pi_0} \delta_0$ and $H_{\Pi_s} \Sigma H_{\Pi_s}$ 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 \( \hat{\theta} \xrightarrow{p} \theta_0 \) instead of the condition that \( \hat{\theta}_s \xrightarrow{p} \theta_0^* \). 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 \( \hat{\theta}_s \), we first adopt a similar strategy in Fan and Li (2001) and Wang and Leng (2007) and show that \( \hat{\theta}_s - \theta_0^* = O_p(n^{-1/2}) \). We then show that \( \hat{\theta}_s \) is approximatively a linear function of \( \hat{\theta} \) as follows

\[
\hat{\theta}_s - \theta_0^* = \hat{H}_{\Pi_s} \hat{H}(\hat{\theta} - \theta_0) + \hat{H}_{\Pi_s} \hat{H}(\theta_0 - \theta_0^*) + o_p(n^{-1/2}), \tag{2.5}
\]

where \( \hat{H}_{\Pi_s} = \Pi_s'(\Pi_s \hat{H}_n \Pi_s')^{-1} \Pi_s \) and \( \hat{H} = H_n(\hat{\theta}) \). 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 \( Z \).
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 \( \mu = \mu(\theta) = \mu(\beta, \gamma) \) be a focus parameter, which is a smooth real-valued function. Notice that if \( \mu \) depends only on \( \gamma \) and the estimator in a model that set \( \gamma = 0 \), then Assumption 1(ii) does not hold. This is because the set \( \Theta \) includes only one point \( \gamma = 0 \), and there is no interior in the set \( \Theta \). Let \( \mu_0 = \mu(\theta_0) = \mu(\beta_0, \delta_0/\sqrt{n}) \)
3.1 The FIC for extremum estimators

be the focus parameter evaluated at $\theta_0$. For the $s$th submodel, $\mu_0$ is estimated by $\hat{\mu}_s = \mu(\hat{\theta}_s)$. Assume that the partial derivatives of $\mu(\theta)$ are continuous in a neighborhood of $\theta^*_0$. Let $D_\theta = (D'_\beta, D'_\gamma)'$ be partial derivatives evaluated at the null points $\theta^*_0$, that is,

$$D_\beta = \left. \frac{\partial \mu(\theta)}{\partial \beta} \right|_{\theta=\theta^*_0} \quad \text{and} \quad D_\gamma = \left. \frac{\partial \mu(\theta)}{\partial \gamma} \right|_{\theta=\theta^*_0}.$$

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

$$\sqrt{n}(\hat{\mu}_s - \mu_0) \xrightarrow{d} \Lambda_s \equiv D'_\theta(H_{\Pi_0}H - I_{p+q})\Pi_0\delta_0 + D'_\theta H_{\Pi_0}HZ \sim N(D'_\theta(H_{\Pi_0}H - I_{p+q})\Pi_0\delta_0, D'_\theta H_{\Pi_0}\Sigma H_{\Pi_0}D_\theta). \quad (3.1)$$

From Corollary 1, a direct calculation yields

$$E(\Lambda_s^2) = D'_\theta(H_{\Pi_0}H - I_{p+q})\Pi_0\delta_0\delta'_0\Pi'_0(H_{\Pi_0}H - I_{p+q})D_\theta$$

$$+ D'_\theta H_{\Pi_0}\Sigma H_{\Pi_0}D_\theta. \quad (3.2)$$

Since $D_\theta$ depends on the focus parameter $\mu$, we can use (3.2) to select a proper submodel depending on the parameter of interest. To use (3.2) for
3.1 The FIC for extremum estimators

model selection, we need to replace the unknown parameters $D\theta$, $H$, $\Sigma$, and $\delta_0$ with the sample analogs. The proposed FIC of the $s$th submodel is defined as

$$
FIC_s = \hat{D}'\theta (\hat{H} - I_{p+q}) \Delta_{0} \hat{\delta}' \Pi'_{s} (\hat{H} - I_{p+q}) \hat{D}_{\theta}
$$

which is an asymptotically unbiased estimator of the mean squared error $E(\Lambda^2)$ in the sense that the mean of the asymptotic distribution of $FIC_s$ equals the mean squared error $E(\Lambda^2)$. Here, $\hat{\delta}'$ 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 $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$ is the extremum estimator from the full model. Define $\hat{D}\theta = \partial\mu(\theta)/\partial\theta|_{\theta=\hat{\theta}}$, where $\hat{\theta} = (\hat{\beta}', 0)'$. Since $\hat{\theta}$ is a consistent estimator of $\theta_0$ by (2.3), it follows that $\hat{D}\theta$ is a consistent estimator of $D\theta$. For the covariance matrix $H$, we can consistently estimate $H$ by the sample analog $\hat{H}$ under Assumption [1]. Similarly, the covariance matrix $\Sigma$ can also be consistently estimated by the sample analog $\hat{\Sigma}$.

We now consider the estimator for the local parameter $\delta_0$. Unlike $D\theta$, $H$, and $\Sigma$, the consistent estimator for $\delta_0$ is not available due to the local asymptotic framework. We can, however, construct an asymptotically un-
3.2 Example: Nonlinear least squares estimator

biased estimator of $\delta_0$ by using the extremum estimator from the full model. The asymptotically unbiased estimator of $\delta_0$ is defined as $\hat{\delta} = \sqrt{n}\hat{\gamma}$. From (2.3) and Assumption 3 we can show that

$$\hat{\delta} - \delta_0 = \sqrt{n}\Pi'_0(\hat{\theta} - \theta_0) \overset{d}{\to} N(0, \Pi'_0\Sigma\Pi_0) .$$

(3.4)

As shown above, $\hat{\delta}$ is an asymptotically unbiased estimator of $\delta_0$. Therefore, the asymptotically unbiased estimator of $\delta_0\delta'_0$ is

$$\hat{\delta}\hat{\delta}' = \delta\delta' - \Pi'_0\Sigma\Pi_0 .$$

(3.5)

3.2 Example: Nonlinear least squares estimator

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

$$y_i = h(x_i, \theta_0) + e_i ,$$

(3.6)

where $\theta_0$ is a vector of unknown parameters, the parametric regression function $h(x_i, \theta)$ is differentiable with respect to $\theta$, and $e_i$ is an unobservable regression error with $E(e_i|x_i) = 0$. If $h(x_i, \theta_0) = x'_i\theta_0$, then we have the classical linear regression model. The NLS estimator $\hat{\theta}$ maximizes the following objective function

$$\hat{Q}_n(\theta) = -\frac{1}{2n} \sum_{i=1}^{n} (y_i - h(x_i, \theta))^2 ,$$

(3.7)

where $1/2$ is a scale factor that has no effect on the asymptotic results. Note that maximizing $\hat{Q}_n(\theta)$ is equivalent to minimizing the sum of squared
3.2 Example: Nonlinear least squares estimator

errors $S_n(\theta) = \sum_{i=1}^{n}(y_i - h(x_i, \theta))^2$. Here the objective function $Q_n(\theta)$ converges to $Q_0(\theta) = E(y_i - h(x_i, \theta))^2/2$. Thus,

$$H_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} h(x_i, \theta) \frac{\partial}{\partial \theta'} h(x_i, \theta) - (y_i - h(x_i, \theta)) \frac{\partial^2}{\partial \theta \partial \theta'} h(x_i, \theta) \right),$$

(3.8)

$$H(\theta) = -E\left( \frac{\partial}{\partial \theta} h(x_i, \theta) \frac{\partial}{\partial \theta'} h(x_i, \theta) \right) + E\left( (y_i - h(x_i, \theta)) \frac{\partial^2}{\partial \theta \partial \theta'} h(x_i, \theta) \right),$$

(3.9)

and

$$\Sigma = E\left( e^2 \frac{\partial}{\partial \theta} h(x_i, \theta_0) \frac{\partial}{\partial \theta'} h(x_i, \theta_0) \right).$$

(3.10)

From (3.6) and (3.9), we have $H = H(\theta_0) = -E\left( \frac{\partial}{\partial \theta} h(x_i, \theta_0) \frac{\partial}{\partial \theta'} h(x_i, \theta_0) \right)$.

By Theorem, it follows that

$$\sqrt{n}(\hat{\theta}_s - \theta_0^*) \overset{d}{\to} H\Pi_s H(Z + \Pi_0 \delta_0) \sim N(H\Pi_s H\Pi_0 \delta_0, V\Pi_s),$$

(3.11)

where $V\Pi_s = H\Pi_s \Sigma H\Pi_s$ and $H\Pi_s = \Pi_s' (\Pi_s\Pi\Pi_s')^{-1} \Pi_s$. In the supplementary materials, we verify the high-level assumptions for the NLS estimator.

Thus, by Corollary, the FIC for the NLS estimator is defined as

$$\text{FIC}_s = \hat{D}_s' (\hat{H}_{\Pi_s} \hat{H} - I_{p+q}) \Pi_0 \delta \hat{\delta}' T\Pi_0' (\hat{H}_{\Pi_s} \hat{H} - I_{p+q}) \hat{D}_s$$

$$+ \hat{D}_s' \hat{V}_{\Pi_s} \hat{D}_s,$$

(3.12)

where $\hat{D}_s$, $\hat{H}$, and $\hat{\Sigma}$ are the sample analogs of $D_\theta$, $H$, and $\Sigma$, and $\hat{\delta} \hat{\delta}'$ is the asymptotically unbiased estimator of $\delta_0 \delta_0'$. 
3.3 Plug-in averaging method

When the error term \( e_i \) is homoskedastic, i.e., \( E(e_i^2|x_i) = \sigma^2 \), we have \( \Sigma = -\sigma^2H \), and the covariance matrix \( V_{\Pi_s} \) is simplified as \( -\sigma^2H_{\Pi_s} \). In this case, the FIC for the NLS estimator is defined as

\[
FIC_s = \hat{D}'_{\theta}(\hat{H}_{\Pi_s}\hat{H} - I_{p+q})\Pi_0\hat{\delta}\Pi_0'(\hat{H}_{\Pi_s}\hat{H} - I_{p+q})\hat{D}_{\theta} - \hat{\sigma}^2\hat{D}'_{\theta}\hat{H}_{\Pi_s}\hat{D}_{\theta},
\]

where \( \hat{\sigma}^2 \) is the sample analog of \( \sigma^2 \).

### 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 \( s \)th submodel and \( w = (w_1, \ldots, w_S)' \) be a weight vector belonging to the weight set \( W = \{ w \in [0, 1]^S : \sum_{s=1}^S w_s = 1 \} \). That is, the weight vector lies in the unit simplex in \( \mathbb{R}^S \). The model averaging estimator of \( \mu_0 \) is defined as

\[
\hat{\mu}(w) = \sum_{s=1}^S w_s\hat{\mu}_s.
\]

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
3.3 Plug-in averaging method

FIC\(_s\) and gives other models zero weights. Thus, the weight function of the FIC is \(\hat{w}_s = 1\{\text{FIC}_s = \min(\text{FIC}_1, \text{FIC}_2, \ldots, \text{FIC}_S)\}\), where \(1\{\cdot\}\) 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

\[
\sqrt{n}(\hat{\mu}(w) - \mu_0) \xrightarrow{d} N(D_0'B(w)\Pi_0\delta_0, V(w)),
\]

where

\[
B(w) = \sum_{s=1}^{S} w_s(H_{\Pi_s}H - I_{p+q})
\]

and

\[
V(w) = \sum_{s=1}^{S} w_s^2 D_0'\Sigma H_{\Pi_s} D_0 + 2 \sum_{s \neq r} w_s w_r D_0'\Sigma H_{\Pi_s} H_{\Pi_r} D_0.
\]
3.3 Plug-in averaging method

Theorem 2 shows the asymptotic normality of the averaging estimator with fixed weights, and it implies that \( \hat{\mu}(w) \) is root-\( n \) consistent. The asymptotic bias and variance of the averaging estimator are \( D'_\theta B(w)\Pi_0\delta_0 \) and \( V(w) \), respectively.

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

\[
A(w) = w'\Psi w, \tag{3.16}
\]

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

\[
\Psi_{s,r} = D'_\theta \left( B_s \Pi_0 \delta_0 \delta_0' \Pi'_r B_r' + H\Pi_\Sigma H\Pi_r \right) D_\theta, \tag{3.17}
\]

and \( B_s = H\Pi_\Sigma H - I_{p+q} \). We then define the optimal fixed-weight vector as

\[
w_{opt} = \arg\min_{w \in \mathcal{W}} w'\Psi w, \tag{3.18}
\]

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

The optimal weight vector, however, is infeasible, since \( \Psi \) is unknown. A feasible version of \( w_{opt} \) could be obtained by replacing the unknown parameters in \( \Psi \) with their sample analogs. As we discussed in Section 3.1, the unknown parameters \( D_\theta \), \( H \), and \( \Sigma \) can be consistently estimated by the sample analogs. Notice that a consistent estimator for \( \delta_0 \) is not available.
3.3 Plug-in averaging method

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 $A(w)$ as follows

$$\hat{A}(w) = w'\hat{\Psi}w,$$  

(3.19)

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

$$\hat{\Psi}_{s,r} = D'_\theta \left( B_\delta \Pi_\delta' \hat{\delta}' \Pi'_\delta B'_\delta + H_{\Pi} \hat{\Sigma} H_{\Pi} \right) D_{\theta},$$  

(3.20)

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

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

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

(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 $\hat{w}$ to construct a plug-in estimator of $\mu_0$ as follows

$$\hat{\mu}(\hat{w}) = \sum_{s=1}^{S} \hat{w}_s \hat{\mu}_s.$$  

(3.22)

As mentioned by Hjort and Claeskens (2003) and Liu (2015), we can also estimate $A(w)$ by inserting $\hat{\delta}$ for $\delta_0$ directly. Thus, the alternative
3.4 Asymptotic behavior of the FIC and plug-in averaging method

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

$$\tilde{\Psi}_{s,r} = D_\theta' \left( \hat{B}_s \Pi_0 \hat{\delta} \hat{\delta}' \Pi_0' \hat{B}_r' + \hat{H}_\Pi \hat{\Sigma} \hat{H}_\Pi \right) D_\theta. \tag{3.23}$$

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 $\hat{\mu}(\hat{w})$. As mentioned in the previous section, $\hat{D}_\theta$, $\hat{H}$, and $\hat{\Sigma}$ are consistent estimators for $D_\theta$, $H$, and $\Sigma$, respectively, and $\hat{\delta} \overset{d}{\to} Z_\delta \sim N(\delta_0, \Pi_0' H^{-1} \Sigma H^{-1} \Pi_0)$ by (3.4). Therefore, it follows that

$$\text{FIC}_s \overset{d}{\to} D_\theta' \left( H_\Pi H - I_{p+q} \right) \Pi_0 \left( Z_\delta Z_\delta' - \Pi_0' H^{-1} \Sigma H^{-1} \Pi_0 \right) \Pi_0' \left( H_\Pi H - I_{p+q} \right)' D_\theta$$

$$+ D_\theta' H_\Pi \Sigma H_\Pi D_\theta. \tag{3.24}$$

This result shows that the proposed FIC defined in (3.3) will not converge in probability to the AMSE of $\hat{\mu}_s$, although $\text{FIC}_s$ is an asymptotically unbiased estimator of $E(\Lambda_s^2)$ 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 $\hat{\mu}(\hat{w})$. 

Statistica Sinica: Newly accepted Paper

(accepted author-version subject to English editing)
3.4 Asymptotic behavior of the FIC and plug-in averaging method

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

$$\hat{\mathbf{w}} \overset{d}{\to} \mathbf{w}^\infty = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathbf{w}'\Psi^\infty\mathbf{w} \quad (3.25)$$

and

$$\sqrt{n}(\hat{\mu}(\hat{\mathbf{w}}) - \mu_0) \overset{d}{\to} \sum_{s=1}^{S} w_s^\infty \Lambda_s, \quad (3.26)$$

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

$$\Psi_{s,r}^\infty = D'_\theta \left( B_s \Pi_0 (Z_\delta Z_\delta' - \Pi'_0 \Lambda^{-1} S \Lambda^{-1} \Pi_0) \Pi'_0 B'_r + H_{\Pi_s} \Sigma H_{\Pi_r} \right) D_{\theta}. \quad (3.27)$$

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 $\hat{\delta}\hat{\delta}'$ 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...
3.4 Asymptotic behavior of the FIC and plug-in averaging method

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 $w' \Psi^\infty w$ is a convex minimization problem when
$w' \Psi^\infty w$ is quadratic, $\Psi^\infty$ is positive definite, and $\mathcal{W}$ is convex. Hence,
$w' \Psi^\infty w$ 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

$$
\tilde{\Psi}_{s,r} \overset{d}{\to} D_\theta' \left( B_s \Pi_0 Z_{\delta} Z_{\delta}' \Pi_0' B_r' + H_{\Pi_0} \Sigma H_{\Pi_1} \right) D_\theta.
$$

(3.28)

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

Remark 4. Using Theorem 2, we can easily apply the plug-in averaging
method to different model setups, and then obtain the asymptotic dis-
tribution of the plug-in estimator based on Corollary 2. For example, if
$Q_n(\cdot)$ is the sum of squared errors with $h(x_i, \theta_0) = x_i' \theta_0$, then Corollary 2
corresponds to Theorem 3 of Liu (2015). Or, if $Q_n(\cdot)$ 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 $\mu$ 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 $h(\cdot) = \exp(x_0'\theta)$, $p = 1$, $q = 2$, $M = 3$, $\delta_0 = d(1.5, 1.25)'$, and $d$ varies on a grid between $-4$ and $4$.

We consider the homoskedastic error and set $\sigma^2 = 1$ and $\Sigma = -\sigma^2 H$, where the diagonal elements of $H$ are $-1$, and off-diagonal elements are $-0.5$. The focus parameter is $\mu = \theta_1$, and $D_{\theta} = (1, 0, 0)'$ 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-
4.1 AMSE comparison

mal weights \( w_{\text{opt}} \) 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
4.1 AMSE comparison

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
4.2 Finite sample performance

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:

\[
y_i = \exp(x_i' \beta + z_i' \gamma) + e_i, \tag{4.1}
\]

where \( x_i = (x_{i1}, \ldots, x_{pi})' \sim iid \) Uniform\((-1, 1)\) and \( z_i = (z_{i1}, \ldots, z_{qi})' \sim iid \) Uniform\((-1, 1)\). The error term is generated by \( e_i = \sigma_i \epsilon_i \), where \( \epsilon_i \) 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 \( \beta = (\beta_1, \ldots, \beta_p)' \) be the must-have parameters and \( \gamma = (\gamma_1, \ldots, \gamma_q)' \)
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 \( \gamma_k = \ldots \)
4.2 Finite sample performance

\[ n^{-1/2}((q - k + 1)/q) \] 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 \( \tilde{\sigma}_s^2 = \frac{1}{n} \sum_{i=1}^{n} \tilde{e}_{si}^2 \), where \( \tilde{e}_{si} \) is the NLS residual from the model \( s \). The AIC of the \( s \)th model is

\[ AIC_s = n\log(\tilde{\sigma}_s^2) + 2(p + q_s), \]

where \( p + q_s \) is the number of parameters in the model \( s \), while the BIC of the \( s \)th model is

\[ BIC_s = n\log(\tilde{\sigma}_s^2) + \log(n)(p + q_s). \]

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

\[ \hat{w}_s = \exp(-\frac{1}{2}AIC_s)/\sum_{r=1}^{S} \exp(-\frac{1}{2}AIC_r). \]

The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors, and the SBIC weight is

\[ \hat{w}_s = \exp(-\frac{1}{2}BIC_s)/\sum_{r=1}^{S} \exp(-\frac{1}{2}BIC_r). \]

Our focus parameter is \( \mu = \beta_p \), which is the last element of the must-
4.2 Finite sample performance

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 $\hat{\mu}$. The MSE is calculated by the average of $(\hat{\mu} - \mu)^2$ 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.
4.2 Finite sample performance

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-
4.3 Real data analysis

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-
4.3 Real data analysis

Table 1: Estimation results

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

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

The simulation scenario 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 \(\tau\), we construct the samples as
\[
y^*_i = \exp(x'_i\hat{\beta}_\tau + z'_i\hat{\gamma}_\tau) + e^*_i,
\]
where \(z'_i\) are the potentially relevant regressors included in the submodel selected by \(\tau\), \(\hat{\beta}_\tau\) and \(\hat{\gamma}_\tau\) are the estimated coefficients from the submodel selected by \(\tau\), and \(e^*_i\) is an i.i.d. random error. The random error is generated by
\[
e^*_i = \hat{\sigma}_\tau \epsilon_i, \text{ where } \epsilon_i \sim iid \text{ Lognormal}(0, 1) \text{ and } \hat{\sigma}_\tau \text{ is the standard error estimated from the submodel selected by } \tau.\]
We then apply the model selection and model averaging methods to the samples \(\{y^*_i, x_i, z_i\}\) and estimate the focus parameter \(\mu\), that is, the coefficient of Education. Notice that the true value of \(\mu\) is known for each choice of \(\tau\). From Table 1, the true values of \(\mu\) 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 \(\hat{\mu}\) 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

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

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


REFERENCES


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


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