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

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

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

[2]*Institute of Economics, Academia Sinica*

Supplementary Material

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

S1. Proofs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Assumption 2, it follows that

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

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

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

Therefore, we have

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

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

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

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

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

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

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

This completes the proof. ■

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

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

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

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

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

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

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

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

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

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

This completes the proof. ■

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

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

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

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

This completes the proof. ■

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

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

discussion.

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

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

Therefore, Assumption 2 holds in this example if

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

and

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

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

S3. Additional examples

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

S3.1 Maximum likelihood estimator

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

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

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

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

and

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

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

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

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

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

Thus, the FIC for the MLE estimator is defined as

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

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

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

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

S3.2 Generalized method of moments estimator

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

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

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

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

By Theorem 1, it follows that

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

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

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

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

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

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

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

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

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

S3.3 Minimum distance estimator

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

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

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

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

S4. Additional numerical results

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

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

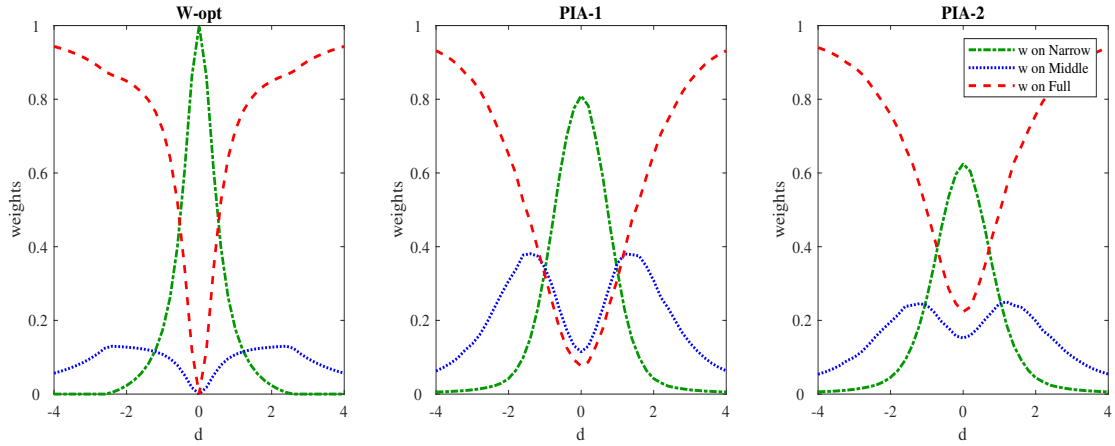

Figure 1: Model weights placed on each submodels

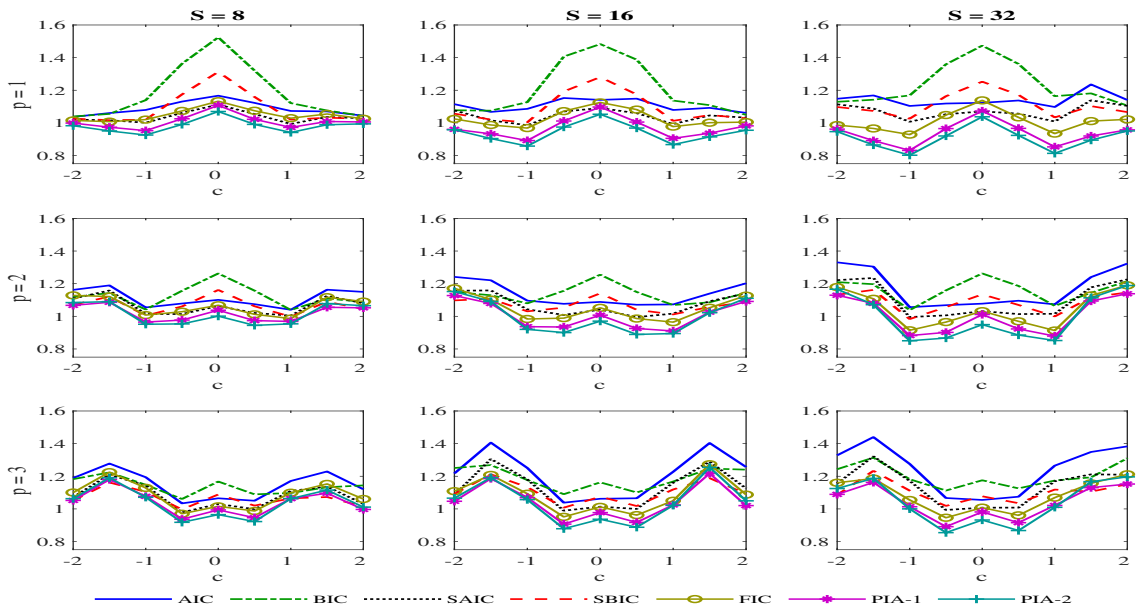

Figure 2: Relative MSE, heteroskedastic errors, $n = 100$.

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

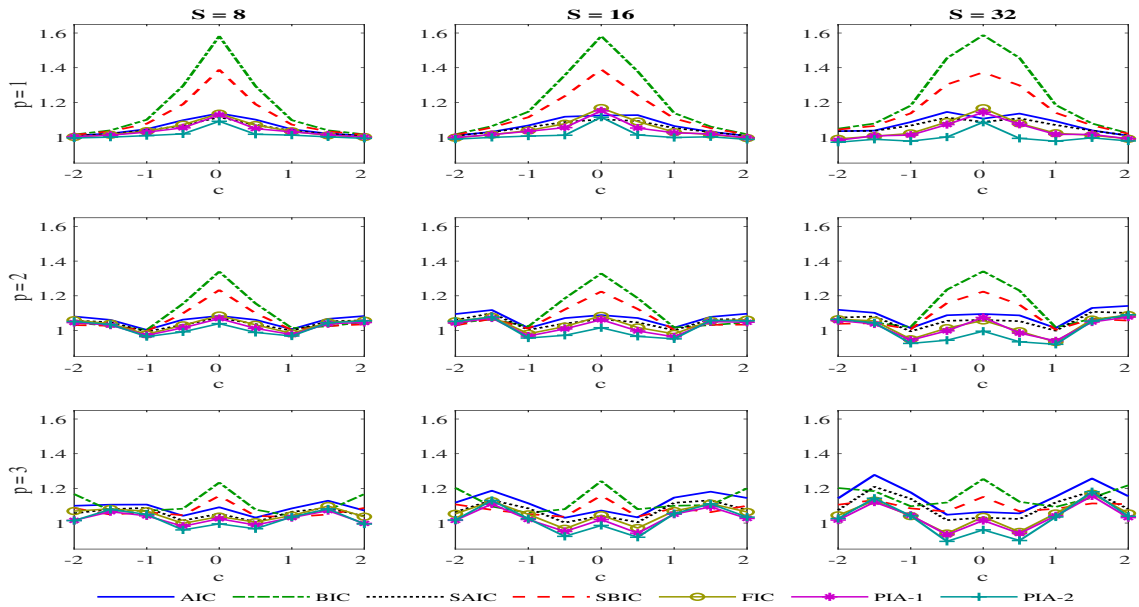

Figure 3: Relative MSE, heteroskedastic errors, $n = 250$.

S5. Post-averaging inference

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This completes the proof. ■

References

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

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

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

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

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

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

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

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

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

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

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

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

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