Supplementary material for "Optimal averaging estimation for density functions"

Peng Lin[a,*], Jun Liao[b,*], Zudi Lu[c,d], Kang You[e] and Guohua Zou[e,¶]

[a] *Shandong University of Technology*

[b] *Renmin University of China*

[c] *University of Southampton*

[d] *City University of Hong Kong*

[e] *Capital Normal University*

This supplementary material contains the derivation of $tr\left(\Sigma_{12}\right)$, Lemma 1, the proofs of all theorems, some illustrating examples, the explanations on the technical conditions, the discussion on the different density aggregation methods and the numerical results (Figures 1–8).

Appendices

A.1 Derivation of $tr\left(\Sigma_{12}\right)$

To derive $tr\left(\Sigma_{12}\right)$, we recall that

$$\widehat{U}(\theta_0) \;=\; \left(\widehat{U}_1(\theta_{10})', ..., \widehat{U}_M(\theta_{M0})'\right)'$$

*Co-first authors.

¶Corresponding author: Guohua Zou. Email: *ghzou@amss.ac.cn*.

$$= \left(n^{-1}\sum_{i=1}^{n} u_1(X_i,\theta_{10})', ..., n^{-1}\sum_{i=1}^{n} u_M(X_i,\theta_{M0})' \right)', \quad \text{(A.1)}$$

and denote

$$\begin{aligned}\widehat{U}(\theta_0,w) &= \left(n^{-1}\sum_{i=1}^{n} \frac{\partial \log f(X_i,\theta_0,w)}{\partial \theta_{10}'}, ..., n^{-1}\sum_{i=1}^{n} \frac{\partial \log f(X_i,\theta_0,w)}{\partial \theta_{M0}'} \right)' \\ &\equiv \left(\widehat{U}_1(\theta_0,w)', ..., \widehat{U}_M(\theta_0,w)' \right)'. \quad \text{(A.2)}\end{aligned}$$

Then from (3.5), it follows that

$$\begin{aligned}&\mathrm{Cov}\left(\widehat{U}_1(\theta_{10}), \widehat{U}_1(\theta_0,w)\right) \\ =\ & E\left(\widehat{U}_1(\theta_{10})\widehat{U}_1(\theta_0,w)'\right) \\ =\ & E\left\{ n^{-1}\sum_{i=1}^{n} \frac{\partial \log f_1(X_i,\theta_{10})}{\partial \theta_{10}} \left(n^{-1}\sum_{i=1}^{n} \frac{1}{f(X_i,\theta_0,w)} \frac{w_1 \partial f_1(X_i,\theta_{10})}{\partial \theta_{10}} \right)' \right\} \\ =\ & E\left\{ n^{-1}\sum_{i=1}^{n} \frac{\partial \log f_1(X_i,\theta_{10})}{\partial \theta_{10}} \left(n^{-1}\sum_{i=1}^{n} \frac{w_1 f_1(X_i,\theta_{10})}{f(X_i,\theta_0,w)} \frac{\partial \log f_1(X_i,\theta_{10})}{\partial \theta_{10}} \right)' \right\} \\ =\ & n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left\{ \frac{\partial \log f_1(X_i,\theta_{10})}{\partial \theta_{10}} \left(\frac{w_1 f_1(X_j,\theta_{10})}{f(X_j,\theta_0,w)} \frac{\partial \log f_1(X_j,\theta_{10})}{\partial \theta_{10}} \right)' \right\} \\ =\ & n^{-1} E\left\{ \frac{w_1 f_1(X,\theta_{10})}{f(X,\theta_0,w)} \frac{\partial \log f_1(X,\theta_{10})}{\partial \theta_{10}} \left(\frac{\partial \log f_1(X,\theta_{10})}{\partial \theta_{10}} \right)' \right\}.\end{aligned}$$

Similarly, for any $1 \le m \le M$, we have

$$\begin{aligned}&\mathrm{Cov}\left(\widehat{U}_m(\theta_{m0}), \widehat{U}_m(\theta_0,w)\right) \\ =\ & n^{-1} E\left\{ \frac{w_m f_m(X,\theta_{m0})}{f(X,\theta_0,w)} \frac{\partial \log f_m(X,\theta_{m0})}{\partial \theta_{m0}} \left(\frac{\partial \log f_m(X,\theta_{m0})}{\partial \theta_{m0}} \right)' \right\}.\end{aligned}$$

This, together with (3.15), shows that

$$tr\left(\Sigma_{12}\right)$$

$$
\begin{aligned}
&= tr\left\{\mathrm{Cov}\left(\sqrt{n}Z_1, \sqrt{n}Z_2\right)\right\} \\
&= tr\left\{n\mathrm{Cov}\left(J^{-1}(\theta_0)\widehat{U}(\theta_0), \widehat{U}(\theta_0, w)\right)\right\} \\
&= tr\left\{nJ^{-1}(\theta_0)\mathrm{Cov}\left(\widehat{U}(\theta_0), \widehat{U}(\theta_0, w)\right)\right\} \\
&= n\sum_{m=1}^{M} tr\left\{J_m^{-1}(\theta_{m0})\mathrm{Cov}\left(\widehat{U}_m(\theta_{m0}), \widehat{U}_m(\theta_0, w)\right)\right\} \\
&= \sum_{m=1}^{M} tr\left[J_m^{-1}(\theta_{m0})E\left\{\frac{w_m f_m(X,\theta_{m0})}{f(X,\theta_0,w)}\frac{\partial \log f_m(X,\theta_{m0})}{\partial\theta_{m0}}\right.\right. \\
&\qquad\qquad \left.\left.\cdot\left(\frac{\partial \log f_m(X,\theta_{m0})}{\partial\theta_{m0}}\right)'\right\}\right] \\
&\approx \sum_{m=1}^{M} tr\left[J_m^{-1}(\theta_{m0})E\left\{\frac{w_m f_m(X,\theta_{m0})}{f(X,\theta_0,w)}\right\}E\left\{u_m(X,\theta_{m0})u_m(X,\theta_{m0})'\right\}\right] \\
&= \sum_{m=1}^{M} E\left\{\frac{w_m f_m(X,\theta_{m0})}{f(X,\theta_0,w)}\right\} tr\left\{J_m^{-1}(\theta_{m0})K_m(\theta_{m0})\right\}, \qquad \text{(A.3)}
\end{aligned}
$$

which implies (3.17).

A.2 A lemma for $\widehat{\theta}$

First, the Taylor expansion of $f_m(x, \theta_{m0} + t)$ at θ_{m0} can be written as

$$-\log f_m(x, \theta_{m0} + t) + \log f_m(x, \theta_{m0}) = -u_m(x, \theta_{m0})'t + R(x, t),$$

where $R(x,t) = -\frac{1}{2}t'I_m(x,\theta_{m0})t + o(\|t\|^2)$. Then, we list the associated conditions for Lemma 1.

Condition 1. $u_m(X, \theta_{m0})$ has a finite covariance matrix.

Condition 2. As $t \to 0$, $ER(X,t) = \frac{1}{2}t' J_m(\theta_{m0})t + o(\|t\|^2)$, and $ER(X,t)^2 = o(\|t\|^2)$.

Conditions 1 and 2 are the same as those for Theorem 2.1 of Hjort and Pollard (1993).

Lemma 1. *Under Conditions 1 and 2, for fixed $p_{\max}$ and M, we have*

$$\hat{\theta} = \theta_0 + J^{-1}(\theta_0)\widehat{U}(\theta_0) + o_p(n^{-1/2}), \tag{A.4}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} J^{-1}(\theta_0)U \sim N_k(0, J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0)), \tag{A.5}$$

where the definitions of $J^{-1}(\theta_0)$, $\widehat{U}(\theta_0)$ and $K(\theta_0)$ can be found in Section 3.1.

Proof of Lemma 1

Under Conditions 1 and 2, we yield from Theorem 2.1 of Hjort and Pollard (1993) that

$$\hat{\theta}_m = \theta_{m0} + J_m^{-1}(\theta_{m0})\widehat{U}_m(\theta_{m0}) + o_p(n^{-1/2}). \tag{A.6}$$

Using the Central Limit Theorem (CLT) for independent and identically distributed (i.i.d) variables, we have that $\sqrt{n}\widehat{U}_m(\theta_{m0}) \xrightarrow{d} U_m \sim N_{p_m}(0, K_m(\theta_{m0}))$,

the p_m-dimensional normal random vector with mean 0 and covariance matrix $K_m(\theta_{m0})$, where $K_m(\theta_{m0}) = \mathrm{Var}\left(u_m(X, \theta_{m0})\right)$. Further, together with (A.6), it is seen that

$$\sqrt{n}(\hat{\theta}_m - \theta_{m0}) \xrightarrow{d} J_m^{-1}(\theta_{m0})U_m \sim N_{p_m}(0, J_m^{-1}(\theta_{m0})K_m(\theta_{m0})J_m^{-1}(\theta_{m0})) \tag{A.7}$$

Meanwhile, we yield from (A.6) that

$$\hat{\theta} = \theta_0 + J^{-1}(\theta_0)\widehat{U}(\theta_0) + o_p(n^{-1/2}). \tag{A.8}$$

Clearly, $\sqrt{n}\widehat{U}(\theta_0) \xrightarrow{d} U \sim N_k(0, K(\theta_0))$. Then, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} J^{-1}(\theta_0)U \sim N_k(0, J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0)). \tag{A.9}$$

This completes the proof of Lemma 1.

A.3 Proof of Theorem 1

Let

$$\widetilde{C}^{\mathrm{DMA}}(w) = n^{-1}C^{\mathrm{DMA}}(w) + \int g(x)\log g(x)dx. \tag{A.10}$$

Apparently, $\mathrm{argmin}_{w\in\mathcal{W}}\widetilde{C}^{\mathrm{DMA}}(w)$ is equivalent to $\mathrm{argmin}_{w\in\mathcal{W}}C^{\mathrm{DMA}}(w)$.

From the proof of Theorem 1 of Wan et al. (2010), we see that Theorem 1 is valid if the following are satisfied:

$$\sup_{w\in\mathcal{W}} \frac{\left|\widetilde{C}^{\mathrm{DMA}}(w) - \mathrm{KL}^*(w)\right|}{\mathrm{KL}^*(w)} = o_p(1), \tag{A.11}$$

and

$$\sup_{w\in\mathcal{W}}\frac{|\mathrm{KL}(w)-\mathrm{KL}^*(w)|}{\mathrm{KL}^*(w)}=o_p(1). \tag{A.12}$$

We first prove (A.11). We observe that

$$\begin{aligned}
&\left|\widetilde{C}^{\mathrm{DMA}}(w)-\mathrm{KL}^*(w)\right| \\
\leq\ &\left|n^{-1}\sum_{i=1}^{n}\log f(X_i,\widehat{\theta},w)-\int g(x)\log f(x,\theta_0,w)dx\right| \\
&+\sum_{m=1}^{M}n^{-2}\sum_{i=1}^{n}\left\{\frac{w_m f_m(X_i,\widehat{\theta}_m)}{f(X_i,\widehat{\theta},w)}\right\}tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\} \\
\equiv\ &\Delta_1(w)+\Delta_2(w).
\end{aligned} \tag{A.13}$$

Hence, to show (A.11), we need only to verify that

$$\sup_{w\in\mathcal{W}}\frac{\Delta_1(w)}{\mathrm{KL}^*(w)}=o_p(1), \tag{A.14}$$

and

$$\sup_{w\in\mathcal{W}}\frac{\Delta_2(w)}{\mathrm{KL}^*(w)}=o_p(1). \tag{A.15}$$

we are currently in a position to prove (A.14). It is seen that

$$\begin{aligned}
\Delta_1(w)\ \leq\ &\left|n^{-1}\sum_{i=1}^{n}\log f(X_i,\widehat{\theta},w)-n^{-1}\sum_{i=1}^{n}\log f(X_i,\theta_0,w)\right| \\
&+\left|n^{-1}\sum_{i=1}^{n}\log f(X_i,\theta_0,w)-\int g(x)\log f(x,\theta_0,w)dx\right| \\
\equiv\ &\Delta_{1,1}(w)+\Delta_{1,2}(w).
\end{aligned} \tag{A.16}$$

We will derive the stochastic orders of the two terms in (A.16). For $\Delta_{1,1}(w)$, by Taylor's expansion, we have

$$\begin{aligned}\Delta_{1,1}(w) =& \left[n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial \log f(X_i,\theta,w)}{\partial\theta'}|_{\theta=\theta_0}\right\}\left(\widehat{\theta}-\theta_0\right)\right] \\ &+\frac{1}{2}\left(\widehat{\theta}-\theta_0\right)'\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial^2 \log f(X_i,\theta,w)}{\partial\theta\theta'}|_{\theta=\widetilde{\theta}}\right\}\right]\left(\widehat{\theta}-\theta_0\right) \\ \equiv& \Delta_{1,1}^{(1)}(w)+\Delta_{1,1}^{(2)}(w),\end{aligned}$$

where $\widetilde{\theta}=(\widetilde{\theta}_1',...,\widetilde{\theta}_M')'$ is a random k-dimensional vector with $\widetilde{\theta}_m$ lying between $\widehat{\theta}_m$ and θ_{m0}. It follows from (A.7) that

$$\left\|\widehat{\theta}_m-\theta_{m0}\right\|=O_p(n^{-1/2}). \tag{A.17}$$

From (A.17), Conditions (C.1) and (C.2) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}|\Delta_{1,1}^{(1)}(w)| &= \left|\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\frac{w_m f_m(X_i,\theta_{m0})}{f(X_i,\theta_0,w)}\frac{\partial \log f_m(X_i,\theta_m)}{\partial\theta_m'}|_{\theta_m=\theta_{m0}}(\widehat{\theta}_m-\theta_{m0})\right| \\ &\leq \max_{1\leq m\leq M}\|\widehat{\theta}_m-\theta_{m0}\|\frac{1}{n}\max_{1\leq m\leq M}\sum_{i=1}^{n}\left\|\frac{\partial \log f_m(X_i,\theta_m)}{\partial\theta_m}|_{\theta_m=\theta_{m0}}\right\| \\ &= O_p(n^{-1/2}).\end{aligned} \tag{A.18}$$

By (A.5), we see that

$$\left\|\widehat{\theta}-\theta_0\right\|=O_p(n^{-1/2}). \tag{A.19}$$

This and Condition (C.3) imply that

$$\sup_{w\in\mathcal{W}} \lambda_{\max}\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial^2 \log f(X_i,\theta,w)}{\partial\theta\theta'}|_{\theta=\widetilde{\theta}}\right\}\right] = O_p(1). \tag{A.20}$$

Combining (A.19) and (A.20), we have

$$\begin{aligned}\sup_{w\in\mathcal{W}}|\Delta_{1,1}^{(2)}(w)| &\leq \sup_{w\in\mathcal{W}}\frac{1}{2}\lambda_{\max}\left[n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial^2 \log f(X_i,\theta,w)}{\partial\theta\theta'}|_{\theta=\widetilde{\theta}}\right\}\right]\left\|\widehat{\theta}-\theta_0\right\|^2 \\ &= O_p(n^{-1}). \end{aligned}\tag{A.21}$$

So, by (A.18) and (A.21), it is seen that

$$\sup_{w\in\mathcal{W}}|\Delta_{1,1}(w)| = O_p(n^{-1/2}). \tag{A.22}$$

We now turn our attention to $\Delta_{1,2}(w)$. Let

$$H_n(w) \equiv n^{-\frac{1+\delta}{2}}\sum_{i=1}^{n}\{\log f(X_i,\theta_0,w) - \mathrm{E}\left[\log f(X_i,\theta_0,w)\right]\}$$

for some $\delta > 0$, and we next show that

$$\sup_{w\in\mathcal{W}} H_n(w) = o_p(1).$$

Denote $h_n = 1/(n^{\frac{1-\delta}{2}}\log n)$. We construct grids by using the regions, which have the form $\mathcal{W}_j = \left\{w : |w - w_{(j)}|_1 \leq h_n\right\}$ with $w_{(j)} = (w_{(j)1},\cdots,w_{(j)M})^\top$ and $|w - w_{(j)}|_1 = \sum_{m=1}^{M}|w_m - w_{(j)m}|$. By selecting $w_{(j)}$ to lay on grids, $\mathcal{W}$ can be covered with $N = O\left(1/h_n^{M-1}\right)$ regions $\mathcal{W}_j$, $j = 1,...,N$. It is clear that

$$\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j}\left|H_n(w) - H_n(w_{(j)})\right|$$

$$\leq \max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j} n^{-\frac{1+\delta}{2}}\sum_{i=1}^{n}\left|\log f(X_i,\theta_0,w)-\log f(X_i,\theta_0,w_{(j)})\right|$$
$$+\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j} n^{-\frac{1+\delta}{2}}\sum_{i=1}^{n}\left|\mathrm{E}\left[\log f(X_i,\theta_0,w)\right]-\mathrm{E}\left[\log f(X_i,\theta_0,w_{(j)})\right]\right|$$
$$\equiv (I)+(II).$$

By Condition (C.1) and Taylor's expansion, it is seen that

$$(I)=\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j} n^{-\frac{1+\delta}{2}}\sum_{i=1}^{n}\left|f^{-1}(X_i,\theta_0,\widetilde{w})\frac{\partial f(X_i,\theta_0,w)}{\partial w'}|_{w=\widetilde{w}}(w-w_{(j)})\right|$$
$$\leq \max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j} n^{-\frac{1+\delta}{2}}\sum_{i=1}^{n} f^{-1}(X_i,\theta_0,\widetilde{w})\sum_{m=1}^{M}|w_m-w_{(j)m}|f_m(X_i,\theta_{m0})$$
$$\leq Cn^{-\frac{1+\delta}{2}}\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j}\sum_{i=1}^{n}\sum_{m=1}^{M}|w_m-w_{(j)m}|$$
$$=O_p(n^{-\frac{1+\delta}{2}}nh_n)=o_p(1),$$

where $\widetilde{w}\in\mathcal{W}$ lies between w and $w_{(j)}$. Similarly, we can prove that $(II)=o(1)$. Hence, it follows that

$$\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j}\left|H_n(w)-H_n(w_{(j)})\right|=o_p(1),$$

and this implies that

$$\sup_{w\in\mathcal{W}}|H_n(w)|\leq \max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j}|H_n(w)|$$
$$\leq \max_{1\leq j\leq N}|H_n(w_{(j)})|+\max_{1\leq j\leq N}\sup_{w\in\mathcal{W}_j}\left|H_n(w)-H_n(w_{(j)})\right|$$
$$=\max_{1\leq j\leq N}|H_n(w_{(j)})|+o_p(1).$$

Consequently, to prove $\sup_{w\in\mathcal{W}}|H_n(w)| = o_p(1)$, it suffices to show that $\max_{1\le j\le N}|H_n(w_{(j)})| = o_p(1)$. Let $u_i(w_{(j)}) = n^{\frac{1-\delta}{2}}\left\{\log f(X_i,\theta_0,w_{(j)}) - \mathrm{E}\left[\log f(X_i,\theta_0,w_{(j)})\right]\right\}$. By Condition (C.1), we have $|u_i(w_{(j)})| \le n^{\frac{1-\delta}{2}}C$ and $\mathrm{Var}\{u_i(w_{(j)})\} \le n^{1-\delta}C$. It is clear that $\mathrm{E}\{u_i(w_{(j)})\} = 0$ and $u_i(w_{(j)})$ are mutually independent. Thus, by Boole's and Bernstein's inequalities, for any $\epsilon > 0$, it is seen that

$$
\begin{aligned}
&\mathrm{P}\left(\max_{1\le j\le N}|H_n(w_{(j)})| \ge \epsilon\right)\\
\le& N\max_{1\le j\le N}\mathrm{P}\left(|H_n(w_{(j)})| \ge \epsilon\right)\\
=& N\max_{1\le j\le N}\mathrm{P}\left(\left|\sum_{i=1}^{n}u_i(w_{(j)})\right| \ge n\epsilon\right)\\
\le& 2N\exp\left(-\frac{n^2\epsilon^2}{2nn^{1-\delta}C + 2Cn^{\frac{1-\delta}{2}}n\epsilon/3}\right)\\
=& 2N\exp\left(-\frac{\epsilon^2}{2Cn^{-\delta} + 2Cn^{-\frac{1+\delta}{2}}\epsilon/3}\right)\\
=& o(1), \qquad\qquad (A.23)
\end{aligned}
$$

where the last equality of (A.23) holds because of $(M-1)\log(n^{\frac{1-\delta}{2}}\log n)/n^{\delta} \to 0$. Hence, we obtain $\sup_{w\in\mathcal{W}}|H_n(w)| = o_p(1)$, which implies that

$$
\sup_{w\in\mathcal{W}}|\Delta_{1,2}(w)| = o_p(n^{\frac{\delta-1}{2}}). \qquad (A.24)
$$

According to (A.16), (A.22), (A.24) and Condition (C.6), (A.14) is immediately obtained.

Now we turn to prove (A.15). From Condition (C.5), we yield that

$$\begin{aligned}
&\left|\sum_{m=1}^{M} n^{-2} \sum_{i=1}^{n} \left\{ \frac{w_m f_m(X_i, \theta_{m0})}{f(X_i, \theta_0, w)} \right\} tr\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m) \right\} \right| \\
\leq\ & n^{-1} \max_{1 \leq m \leq M} \left| tr\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m) \right\} \right| \\
\leq\ & n^{-1} \max_{1 \leq m \leq M} \left[\lambda_{\max}\left\{ \widehat{K}_m(\widehat{\theta}_m) \right\} \right] \max_{1 \leq m \leq M} \left[\lambda_{\max}\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \right\} \right] \\
& \cdot \max_{1 \leq m \leq M} \left[\operatorname{rank}\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \right\} \right] \\
\leq\ & n^{-1} p_{\max}, \qquad \text{(A.25)}
\end{aligned}$$

which implies that (A.15) is true under Condition (C.6), and hence (A.11).

Then, we prove (A.12). Observe from (A.19) and Condition (C.4) that

$$\begin{aligned}
|\mathrm{KL}(w) - \mathrm{KL}^*(w)| &= \left| \int g(x) \log \frac{f(x, \widehat{\theta}, w)}{f(x, \theta_0, w)} dx \right| \\
&= \left| \left[\int g(x) \left\{ \frac{\partial \log f(x, \theta, w)}{\partial \theta'} |_{\theta = \widetilde{\theta}} \right\} dx \right] \left(\widehat{\theta} - \theta_0 \right) \right| \\
&= O_p(n^{-1/2}), \qquad \text{(A.26)}
\end{aligned}$$

which, together with Condition (C.6), implies (A.12). This completes the proof of Theorem 1.

A.4 Proof of Theorem 2

First, we note from (A.18) and Conditions (C.7) and (C.9) that, when M and the dimension of $\hat{\theta}_m$ are diverging,

$$\sup_{w \in \mathcal{W}} |\Delta_{1,1}^{(1)}(w)| = O_p(n^{-1/2} p_{\max}). \qquad \text{(A.27)}$$

In addition, it follows from Condition (C.9) that

$$\left\|\widehat{\theta}-\theta_0\right\| = \left(\sum_{m=1}^{M}\|\hat{\theta}_m-\theta_{m0}\|^2\right)^{1/2} = O_p(p_{\max}^{1/2}M^{1/2}n^{-1/2}). \quad (A.28)$$

So, from (A.21), (A.28) and Condition (C.3), we have

$$\sup_{w\in\mathcal{W}}|\Delta_{1,1}^{(2)}(w)| = O_p(p_{\max}Mn^{-1}). \quad (A.29)$$

Note that (A.24) continues to hold, and thus by (A.27) and (A.29), we see that (A.14) still holds under Condition (C.10). (A.15) is also true because of $p_{\max}/(n\xi_n) = o(1)$ by Condition (C.10), and hence (A.11) is obtained. Finally, by (A.26), (A.28) and Condition (C.8), we have

$$\begin{aligned}|\mathrm{KL}(w)-\mathrm{KL}^*(w)| &\leq E\left\|\frac{\partial\log f(X,\theta,w)}{\partial\theta'}|_{\theta=\widetilde{\theta}}\right\|\left\|\widehat{\theta}-\theta_0\right\| \\ &= O_p(p_{\max}Mn^{-1/2}),\end{aligned}$$

which, together with Condition (C.10), implies (A.12). This completes the proof of Theorem 2.

A.5 Proof of Theorem 3

Denote $\epsilon_n = k^{1/2}p_{\max}^{1/2}Mn^{-1/2+\alpha}$. To verify Theorem 3, following Fan and Peng (2004) and Chen et al. (2018), it suffices to show that, there is a constant C_0 such that, for M dimensional vector $u=(u_1,...,u_M)'$,

$$\lim_{n\to\infty}P\left(\inf_{\|u\|=C_0,(w^0+\epsilon_nu)\in\mathcal{W}}C^{\mathrm{DMA}}(w^0+\epsilon_nu)>C^{\mathrm{DMA}}(w^0)\right)=1, \quad (A.30)$$

which means that there exists a minimum $\hat{w}$ in the set $\{w^0 + \epsilon_n u : \|u\| \leq C_0, (w^0 + \epsilon_n u) \in \mathcal{W}\}$ such that $\|\hat{w} - w^0\| = O_p(\epsilon_n)$.

Observe that

$$\begin{aligned}
& C^{\mathrm{DMA}}(w^0 + \epsilon_n u) - C^{\mathrm{DMA}}(w^0) \\
= & -\sum_{i=1}^{n} \log f(X_i, \widehat{\theta}, w^0 + \epsilon_n u) + \sum_{i=1}^{n} \log f(X_i, \widehat{\theta}, w^0) \\
& + \sum_{m=1}^{M} n^{-1} \sum_{i=1}^{n} \left\{ \frac{(w_m^0 + \epsilon_n u_m) f_m(X_i, \theta_{m0})}{f(X_i, \theta_0, w^0 + \epsilon_n u)} \right\} tr\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m) \right\} \\
& - \sum_{m=1}^{M} n^{-1} \sum_{i=1}^{n} \left\{ \frac{w_m^0 f_m(X_i, \theta_{m0})}{f(X_i, \theta_0, w^0)} \right\} tr\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m) \right\} \\
= & -\sum_{i=1}^{n} \log f(X_i, \widehat{\theta}, w^0 + \epsilon_n u) + \sum_{i=1}^{n} \log f(X_i, \widehat{\theta}, w^0) \\
& + \sum_{m=1}^{M} n^{-1} \sum_{i=1}^{n} \left[\frac{\epsilon_n f_m(X_i, \theta_{m0}) \{u_m f(X_i, \theta_0, w^0) - w_m^0 f(X_i, \theta_0, u)\}}{f(X_i, \theta_0, w^0 + \epsilon_n u) f(X_i, \theta_0, w^0)} \right] \\
& \qquad \cdot tr\left\{ \widehat{J}_m^{-1}(\widehat{\theta}_m) \widehat{K}_m(\widehat{\theta}_m) \right\}, \qquad \text{(A.31)}
\end{aligned}$$

where the last term on the right-hand side of (A.31) is denoted by Δ_3.

Recall that $F(X_i, \widehat{\theta}) = (f_1(X_i, \widehat{\theta}_1), ..., f_M(X_i, \widehat{\theta}_M))'$. Note that

$$\frac{\partial \log f(X_i, \widehat{\theta}, w)}{\partial w'} = \frac{F(X_i, \widehat{\theta})'}{f(X_i, \widehat{\theta}, w)}, \qquad \text{(A.32)}$$

and

$$\frac{\partial^2 \log f(X_i, \widehat{\theta}, w)}{\partial w' w} = -\frac{F(X_i, \widehat{\theta}) F(X_i, \widehat{\theta})'}{f(X_i, \widehat{\theta}, w)^2}. \qquad \text{(A.33)}$$

Therefore, we have

$$\begin{aligned}
& -\sum_{i=1}^{n}\log f(X_i,\widehat{\theta},w^0+\epsilon_n u)+\sum_{i=1}^{n}\log f(X_i,\widehat{\theta},w^0) \\
= & -\epsilon_n\sum_{i=1}^{n}\frac{\partial\log f(X_i,\widehat{\theta},w)}{\partial w'}|_{w=w^0}u \\
& -\epsilon_n^2\sum_{i=1}^{n}u'\int_0^1\int_0^1 v\frac{\partial^2\log f(X_i,\widehat{\theta},w)}{\partial w'w}|_{w=w^0+tv\epsilon_n u}dtdvu \\
= & -\epsilon_n\sum_{i=1}^{n}\frac{F(X_i,\widehat{\theta})'}{f(X_i,\widehat{\theta},w)}|_{w=w^0}u \\
& +\epsilon_n^2u'\sum_{i=1}^{n}\int_0^1\int_0^1 v\frac{F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'}{f(X_i,\widehat{\theta},w)^2}|_{w=w^0+tv\epsilon_n u}dtdvu \\
\equiv & \ \Delta_4+\Delta_5, \qquad \text{(A.34)}
\end{aligned}$$

in which

$$\begin{aligned}
\Delta_5 &= \epsilon_n^2u'\sum_{i=1}^{n}\int_0^1\int_0^1 v\frac{F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'}{f(X_i,\widehat{\theta},w)^2}|_{w=w^0+tv\epsilon_n u}dtdvu \\
&\geq C\epsilon_n^2u'\sum_{i=1}^{n}F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'u \\
&> C\kappa_1\epsilon_n^2n\|u\|^2>0 \qquad \text{(A.35)}
\end{aligned}$$

in probability tending to 1 by Condition (C.16). In the following, we will show that $|\Delta_4|$ is dominated by Δ_5 asymptotically.

Clearly, we can write

$$\begin{aligned}
|\Delta_4| &= |\epsilon_n\sum_{i=1}^{n}\frac{F(X_i,\widehat{\theta})'}{f(X_i,\widehat{\theta},w)}|_{w=w^0}u| \\
&\leq \epsilon_n\left|\sum_{i=1}^{n}\left\{\frac{F(X_i,\widehat{\theta})'u}{f(X_i,\widehat{\theta},w^0)}-\frac{F(X_i,\theta_0)'u}{f(X_i,\theta_0,w^0)}\right\}\right|+\epsilon_n\left|\sum_{i=1}^{n}\frac{F(X_i,\theta_0)'u}{f(X_i,\theta_0,w^0)}\right|.
\end{aligned}$$

(A.36)

First, we consider the order of the second term in (A.36). By Lagrange Multiplier Method, we see that to optimize the objective function $\mathrm{KL}^*(w)$ with the condition $\sum_{m=1}^{M} w_m = 1$, it is equivalent to optimize the following objective function:

$$L(\boldsymbol{w}, \lambda) = \mathrm{KL}^*(\boldsymbol{w}) + \lambda \left(\sum_{m=1}^{M} w_m - 1 \right).$$

Since w^0 is an interior point of $\mathcal{W}$, there exists a constant λ^0 such that for $m = 1, ..., M$,

$$\int g(x) \frac{f_m(x, \theta_{m0})}{f(x, \theta_0, w^0)} dx = \lambda^0$$

Since $w^0 + \epsilon_n u \in \mathcal{W}$ and $\sum_{m=1}^{M} w_m^0 = 1$, we obtain that $\sum_{m=1}^{M} u_m = 0$. Hence, it follows that

$$\begin{aligned} \mathrm{E}\left(\frac{F(X, \theta_0)' u}{f(X, \theta_0, w^0)} \right) &= \sum_{m=1}^{M} u_m \mathrm{E}\left(\frac{f_m(X, \theta_{m0})}{f(X, \theta_0, w^0)} \right) \\ &= \lambda^0 \sum_{m=1}^{M} u_m = 0. \end{aligned}$$

So, by Conditions (C.11) and (C.13), it is seen that

$$\begin{aligned} &\mathrm{E}\left(\sum_{i=1}^{n} \frac{F(X_i, \theta_0)' u}{f(X_i, \theta_0, w^0)} \right)^2 \\ =& \sum_{i=1}^{n} \mathrm{E}\left(\frac{F(X_i, \theta_0)' u}{f(X_i, \theta_0, w^0)} \right)^2 \end{aligned}$$

$$
\begin{aligned}
&\leq n\left\{\mathrm{E}(f^{-4}(X_i,\theta_0,w^0))\right\}^{1/2}\left\{\mathrm{E}\|F(X_i,\theta_0)\|^4\right\}^{1/2}\|u\|^2\\
&\leq nM\left\{\mathrm{E}(f^{-4}(X_i,\theta_0,w^0))\right\}^{1/2}\left\{\mathrm{E}\left(\sup_{1\leq m\leq M} f_m^4(X_i,\theta_{m0})\right)\right\}^{1/2}\|u\|^2\\
&= O(nM).
\end{aligned}
$$

It follows that

$$
\epsilon_n\sum_{i=1}^{n}\frac{F(X_i,\theta_0)'u}{f(X_i,\theta_0,w^0)} = O_p(\epsilon_n n^{1/2}M^{1/2}). \tag{A.37}
$$

Now, we deal with the first term in (A.36). We can write

$$
\begin{aligned}
&\sum_{i=1}^{n}\left\{\frac{F(X_i,\widehat{\theta})'u}{f(X_i,\widehat{\theta},w)}-\frac{F(X_i,\theta_0)'u}{f(X_i,\theta_0,w)}\right\}\\
=\;&\sum_{i=1}^{n}\left\{\frac{1}{f(X_i,\widehat{\theta},w)}-\frac{1}{f(X_i,\theta_0,w)}\right\}F(X_i,\theta_0)'u\\
&+\sum_{i=1}^{n}\left\{\frac{1}{f(X_i,\widehat{\theta},w)}-\frac{1}{f(X_i,\theta_0,w)}\right\}\left\{F(X_i,\widehat{\theta})'u-F(X_i,\theta_0)'u\right\}\\
&+\sum_{i=1}^{n}\frac{1}{f(X_i,\theta_0,w)}\left\{F(X_i,\widehat{\theta})'-F(X_i,\theta_0)'\right\}u\\
\equiv\;&\Delta_6+\Delta_7+\Delta_8.
\end{aligned}
\tag{A.38}
$$

We will derive the stochastic order of each term in (A.38). It is seen from Condition (C.11) that

$$
\begin{aligned}
&E\left|F(X_i,\theta_0)'u\right|^4\\
\leq\;&E\left\|F(X_i,\theta_0)\right\|^4\|u\|^4\\
=\;&E\left\{\sum_{m=1}^{M}f_m^2(X_i,\theta_{m0})\right\}^2\|u\|^4
\end{aligned}
$$

$$
\begin{aligned}
&\leq M^2 E\left\{\sup_{1\leq m\leq M} f_m^4(X_i,\theta_{m0})\right\}\|u\|^4 \\
&= O(M^2), \qquad \text{(A.39)}
\end{aligned}
$$

and hence together with Conditions (C.12) and (C.13), we have

$$
\begin{aligned}
& k^{-1/2}E\left\{\sup_{\theta\in\aleph}\left|\frac{F(X_i,\theta_0)'u}{f(X_i,\theta,w)}\right|\sup_{\theta\in\aleph}\left\|\frac{\partial \log f(X_i,\theta,w)}{\partial\theta}\right\|\right\} \\
\leq\ & \left[E\left\{|F(X_i,\theta_0)'u|\sup_{\theta\in\aleph}\left|f^{-1}(X_i,\theta,w)\right|\right\}^2\right]^{1/2} \\
& \cdot\left[E\left\{\sup_{\theta\in\aleph}\left\|k^{-1/2}\frac{\partial \log f(X_i,\theta,w)}{\partial\theta}\right\|\right\}^2\right]^{1/2} \\
\leq\ & \left[E\left|F(X_i,\theta_0)'u\right|^4 E\left\{\sup_{\theta\in\aleph} f^{-4}(X_i,\theta,w)\right\}\right]^{1/4} \\
& \cdot\left[E\left\{\sup_{\theta\in\aleph}\left\|k^{-1/2}\frac{\partial \log f(X_i,\theta,w)}{\partial\theta}\right\|^2\right\}\right]^{1/2} \\
=\ & O(M^{1/2}). \qquad \text{(A.40)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\left|n^{-1}\Delta_6\right| &= \left|n^{-1}\sum_{i=1}^n\left\{-\frac{F(X_i,\theta_0)'u}{f^2(X_i,\theta,w)}\frac{\partial f(X_i,\theta,w)}{\partial\theta'}\right\}|_{\theta=\widetilde{\theta}}\left(\widehat{\theta}-\theta_0\right)\right| \\
&\leq n^{-1}\sum_{i=1}^n\left|\frac{F(X_i,\theta_0)'u}{f(X_i,\widetilde{\theta},w)}\right|\left\|\frac{\partial \log f(X_i,\theta,w)}{\partial\theta}|_{\theta=\widetilde{\theta}}\right\|\left\|\widehat{\theta}-\theta_0\right\| \\
&\leq n^{-1}\sum_{i=1}^n\sup_{\theta\in\aleph}\left|\frac{F(X_i,\theta_0)'u}{f(X_i,\theta,w)}\right|\sup_{\theta\in\aleph}\left\|\frac{\partial \log f(X_i,\theta,w)}{\partial\theta}\right\|\left\|\widehat{\theta}-\theta_0\right\| \\
&= O_p(k^{1/2}p_{\max}^{1/2}Mn^{-1/2}), \qquad \text{(A.41)}
\end{aligned}
$$

where $\widetilde{\theta}$ is a random k-dimensional vector lying between $\widehat{\theta}$ and θ_0, the second

inequality holds when $\widetilde{\theta}$ is in $\aleph$, and the last equality is obtained from (A.28) and (A.40).

Further, let $\bar{\theta}$ be a random k-dimensional vector lying between $\widehat{\theta}$ and θ_0. Then for $\widetilde{\theta}$ and $\bar{\theta}$ both in $\aleph$,

$$\begin{aligned}
\left|n^{-1}\Delta_7\right| &= \left| n^{-1}\sum_{i=1}^{n}\left(\widehat{\theta}-\theta_0\right)'\left\{-\frac{1}{f^2(X_i,\theta,w)}\frac{\partial f(X_i,\theta,w)}{\partial\theta}\right\}|_{\theta=\widetilde{\theta}}\right. \\
&\qquad \left.\cdot\frac{\partial F(X_i,\theta)'u}{\partial\theta'}|_{\theta=\bar{\theta}}\left(\widehat{\theta}-\theta_0\right)\right| \\
&\leq n^{-1}\sum_{i=1}^{n}\sup_{\theta\in\aleph}\left|f^{-1}(X_i,\theta,w)\right|\sup_{\theta\in\aleph}\left\|\frac{\partial\log f(X_i,\theta,w)}{\partial\theta}\right\| \\
&\qquad \cdot\sup_{\theta\in\aleph}\lambda_{\max}\left\{\frac{\partial F(X_i,\theta)}{\partial\theta'}\right\}\left\|\widehat{\theta}-\theta_0\right\|^2\|u\| \\
&= O_p(k^{1/2}M^{1/2}k^{1/2}p_{\max}Mn^{-1}) \\
&= O_p(k^{1/2}p_{\max}^{1/2}Mn^{-1/2}) \qquad \text{(A.42)}
\end{aligned}$$

because of Condition (C.15), (A.28) and

$$\begin{aligned}
& k^{-1/2}E\sup_{\theta\in\aleph}\left|f^{-1}(X_i,\theta,w)\right|\sup_{\theta\in\aleph}\lambda_{\max}\left\{\frac{\partial F(X_i,\theta)}{\partial\theta'}\right\}\sup_{\theta\in\aleph}\left\|\frac{\partial\log f(X_i,\theta,w)}{\partial\theta}\right\| \\
\leq\ & \left[E\sup_{\theta\in\aleph}f^{-4}(X_i,\theta,w)E\sup_{\theta\in\aleph}\left\{\lambda_{\max}\left(\frac{\partial F(X_i,\theta)}{\partial\theta'}\right)\right\}^4\right]^{1/4} \\
& \cdot\left[E\left\{\sup_{\theta\in\aleph}\left\|k^{-1/2}\frac{\partial\log f(X_i,\theta,w)}{\partial\theta}\right\|^2\right\}\right]^{1/2} \\
=\ & O(M^{1/2}k^{1/2})
\end{aligned}$$

by Conditions (C.12), (C.13) and (C.14).

Finally, by Conditions (C.13) and (C.14), we have

$$\begin{aligned} & E\left[\left|f^{-1}(X_i,\theta_0,w)\right| \sup_{\theta\in\aleph} \lambda_{\max}\left\{\frac{\partial F(X_i,\theta)}{\partial\theta'}\right\}\right] \\ \leq\ & \left\{Ef^{-2}(X_i,\theta_0,w)\right\}^{1/2}\left[E\sup_{\theta\in\aleph}\left\{\lambda_{\max}\left(\frac{\partial F(X_i,\theta)}{\partial\theta'}\right)\right\}^4\right]^{1/4} \\ =\ & O(M^{1/2}k^{1/2}), \end{aligned} \tag{A.43}$$

which, together with (A.28), implies that

$$\begin{aligned} \left|n^{-1}\Delta_8\right| &= \left|n^{-1}\sum_{i=1}^n f^{-1}(X_i,\theta_0,w)\frac{\partial F(X_i,\theta)'u}{\partial\theta'}|_{\theta=\bar{\theta}}\left(\widehat{\theta}-\theta_0\right)\right| \\ &\leq n^{-1}\sum_{i=1}^n\left|f^{-1}(X_i,\theta_0,w)\right| \sup_{\theta\in\aleph}\lambda_{\max}\left\{\frac{\partial F(X_i,\theta)}{\partial\theta'}\right\}\left\|\widehat{\theta}-\theta_0\right\|\|u\| \\ &= O_p(k^{1/2}p_{\max}^{1/2}Mn^{-1/2}). \end{aligned} \tag{A.44}$$

Combining (A.38), (A.41), (A.42) and (A.44), we obtain

$$\epsilon_n\left|\sum_{i=1}^n\left\{\frac{F(X_i,\widehat{\theta})'u}{f(X_i,\widehat{\theta},w^0)}-\frac{F(X_i,\theta_0)'u}{f(X_i,\theta_0,w^0)}\right\}\right| = O_p(k^{1/2}p_{\max}^{1/2}Mn^{1/2})\epsilon_n. \tag{A.45}$$

From (A.37) and (A.45), it is readily seen that $|\Delta_4|$ is asymptotically dominated by $|\Delta_5|$ since

$$\frac{k^{1/2}p_{\max}^{1/2}Mn^{1/2}\epsilon_n}{n\epsilon_n^2} = \frac{k^{1/2}p_{\max}^{1/2}Mn^{1/2}}{nk^{1/2}p_{\max}^{1/2}Mn^{-1/2+\alpha}} \to 0, \tag{A.46}$$

and

$$\frac{M^{1/2}n^{1/2}\epsilon_n}{n\epsilon_n^2} \to 0. \tag{A.47}$$

To prove that $|\Delta_3|$ is asymptotically dominated by $|\Delta_5|$, we note from Conditions (C.11) and (C.13) that

$$\begin{aligned}
& E \sup_{1\le m\le M} \left| \frac{f_m(X_i,\theta_{m0})\left\{u_m f(X_i,\theta_0,w^0) - w_m^0 f(X_i,\theta_0,u)\right\}}{f(X_i,\theta_0,w^0+\epsilon_n u) f(X_i,\theta_0,w^0)} \right| \\
\le\ & C\left\{Ef^{-4}(X_i,\theta_0,w^0+\epsilon_n u)\right\}^{1/4}\left\{Ef^{-4}(X_i,\theta_0,w^0)\right\}^{1/4} \\
& \cdot \left[E\left\{f^4(X_i,\theta_0,w^0)\right\} + E\left\{f^4(X_i,\theta_0,u)\right\}\right]^{1/4} \\
& \cdot \left\{E \sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\}^{1/4} \\
\le\ & C\left[E\left\{\sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\} + M^2\|u\|^4 E\left\{\sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\}\right]^{1/4} \\
& \cdot \left\{E \sup_{1\le m\le M} f_m^4(X_i,\theta_{m0})\right\}^{1/4} \\
=\ & O(M^{1/2}),
\end{aligned}$$

which results in

$$\begin{aligned}
& \left| \sum_{m=1}^{M} n^{-1} \sum_{i=1}^{n} \left[\frac{\epsilon_n f_m(X_i,\theta_{m0})\left\{u_m f(X_i,\theta_0,w^0) - w_m^0 f(X_i,\theta_0,u)\right\}}{f(X_i,\theta_0,w^0+\epsilon_n u) f(X_i,\theta_0,w^0)} \right] \right. \\
& \left. \cdot tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\} \right| \\
\le\ & \epsilon_n M n^{-1} \sum_{i=1}^{n} \sup_{1\le m\le M} \left| \frac{f_m(X_i,\theta_{m0})\left\{u_m f(X_i,\theta_0,w^0) - w_m^0 f(X_i,\theta_0,u)\right\}}{f(X_i,\theta_0,w^0+\epsilon_n u) f(X_i,\theta_0,w^0)} \right| \\
& \cdot \sup_{1\le m\le M} \left| tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\} \right| \\
=\ & O_p(M^{3/2}\epsilon_n) \sup_{1\le m\le M} \left| tr\left\{\widehat{J}_m^{-1}(\widehat{\theta}_m)\widehat{K}_m(\widehat{\theta}_m)\right\} \right| \\
=\ & O_p(M^{3/2} p_{\max}\epsilon_n) \qquad \text{(A.48)}
\end{aligned}$$

where the last equality is obtained by Condition (C.5) (see (A.25)). Thus

$|\Delta_3|$ is also asymptotically dominated by $|\Delta_5|$ due to

$$\frac{M^{3/2}p_{\max}\epsilon_n}{n\epsilon_n^2} = \frac{k^{1/2}M^{1/2}p_{\max}^{1/2}n^{-1/2}}{kn^{\alpha}} \to 0 \tag{A.49}$$

under Condition (C.15). This completes the proof of Theorem 3.

A.6 Proof of Theorem 4

Using Condition (C.18) and Theorem 16(a) of Ferguson (1996), we have

$$n^{-1}\sum_{i=1}^{n}\left\{\log \tilde{f}^0(X_i,\rho) - E\left[\log \tilde{f}^0(X_i,\rho)\right]\right\} \xrightarrow{a.s.} 0$$

uniformly in ρ. This, together with Conditions (C.17) and (C.19), implies that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\left\{\log \tilde{f}^{[-i]}(X_i,\rho)\right\} - T(\rho) \\
= \; & n^{-1}\sum_{i=1}^{n}\left\{\log \tilde{f}^{[-i]}(X_i,\rho) - \log \tilde{f}^0(X_i,\rho)\right\} \\
& + n^{-1}\sum_{i=1}^{n}\left\{\log \tilde{f}^0(X_i,\rho) - E\left[\log \tilde{f}^0(X_i,\rho)\right]\right\} \\
& + E\left[\log \tilde{f}^0(X_i,\rho)\right] - T(\rho) \\
\xrightarrow{p} \; & 0
\end{aligned} \tag{A.50}$$

uniformly in ρ. Let $\rho^* = \text{argmin}_{0\le\rho\le 1} T(\rho)$, and recall that $\tilde{\rho} = \text{argmin}_{0\le\rho\le 1}\{-\sum_{i=1}^{n} \log \tilde{f}^{[-i]}(X_i,\rho)\}$ and $\rho_0 = \text{argmin}_{0\le\rho\le 1}\left\{\text{KL}\left(g, \tilde{f}^0(x,\rho)\right)\right\}$. Then, from Theorem 4.1.1 of Amemiya (1985), we have $\tilde{\rho} \xrightarrow{p} \rho^*$ by (A.50), and $\rho_0 \to \rho^*$

by Condition (C.19), which implies $\tilde{\rho} - \rho_0 \xrightarrow{p} 0$. This completes the proof of Theorem 4.

A.7 Illustrating examples

Example 1. To implement DMA, we need to calculate $\widehat{K}_m(\theta_m)$ and $\widehat{J}_m(\theta_m)$ ($\widehat{K}_m(\widehat{\theta}_m)$ and $\widehat{J}_m(\widehat{\theta}_m)$ can be obtained by replacing θ_m in $\widehat{K}_m(\theta_m)$ and $\widehat{J}_m(\theta_m)$ by its estimator $\widehat{\theta}_m$, respectively). Let $\theta_m = (a, b)'$, then for the gamma distribution with the density $b^a/\Gamma(a)x^{a-1}e^{-bx}$ with $x > 0$, we have

$$\begin{aligned} & \widehat{K}_m(\theta_m) \\ = \; & n^{-1}\sum_{i=1}^{n}\begin{pmatrix} (\log(bX_i) - d(a))^2 & (\log(bX_i) - d(a))\,(a/b - X_i) \\ (\log(bX_i) - d(a))\,(a/b - X_i) & (a/b - X_i)^2 \end{pmatrix} \end{aligned}$$

and

$$\widehat{J}_m(\theta_m) = \begin{pmatrix} T(a) & -1/b \\ -1/b & a/b^2 \end{pmatrix},$$

where $d(a)$ and $t(a)$ are the first-order and second-order derivatives of $\log\Gamma(a)$ with respect to a, respectively. For the log-normal distribution with the density $\frac{1}{\sqrt{2\pi}\sigma}x^{-1}e^{-(\log x-\mu)^2/(2\sigma^2)}$ with $x > 0$, we have

$$\begin{aligned} & \widehat{K}_m(\theta_m) \\ = \; & n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \frac{(\log X_i-\mu)^2}{\sigma^4} & \frac{(\log X_i-\mu)}{\sigma^2}\left(-\frac{1}{\sigma} + \frac{(\log X_i-\mu)^2}{\sigma^3}\right) \\ \frac{(\log X_i-\mu)}{\sigma^2}\left(-\frac{1}{\sigma} + \frac{(\log X_i-\mu)^2}{\sigma^3}\right) & \left(-\frac{1}{\sigma} + \frac{(\log X_i-\mu)^2}{\sigma^3}\right)^2 \end{pmatrix}, \end{aligned}$$

and

$$\widehat{J}_m(\theta_m) = -n^{-1}\sum_{i=1}^{n}\begin{pmatrix} -\frac{1}{\sigma^2} & \frac{-2(\log X_i-\mu)}{\sigma^3} \\ \frac{-2(\log X_i-\mu)}{\sigma^3} & \frac{1}{\sigma^2}-\frac{3(\log X_i-\mu)^2}{\sigma^4} \end{pmatrix},$$

where $\theta_m = (\mu, \sigma)'$. Also, substituting $\log X_i$ above by X_i, we can obtain the corresponding matrices $\widehat{K}_m(\theta_m)$ and $\widehat{J}_m(\theta_m)$ for the Normal distribution $N(\mu, \sigma^2)$.

Example 2. Consider the candidate density family:

$$f_m(x, \theta_m) = f_0(x)e^{\sum_{j=1}^{m} a_j\psi_j(x)}/c_m(\theta_m), \tag{A.51}$$

where $1 \le m \le M$, $\theta_m = (a_1, ..., a_m)'$, $f_0(x)$ is some specified density, $\{\psi_j\}$ are the basis functions which satisfy $\int f_0(x)\psi_j(x)\psi_k(x)dx = 1$ for $j = k$ and 0 for $j \neq k$, $c_m(\theta_m) = \int f_0(x)e^{\sum_{j=1}^{m} a_j\psi_j(x)}dx$. Here a larger m means a higher degree of complexity. See Claeskens and Hjort (2008) for more details. This density family has been frequently considered in Claeskens and Hjort (2008) who use AIC to select the best m for estimating the true density. As the authors commented, the resultant estimator may well do better than full-fledged nonparametric estimators in cases where a low-order sum captures the main aspects of an underlying density curve. Here we consider density averaging estimation over the candidate density family (A.51) based on the proposed method DMA. Specifically, let $\varphi_m(x) = (\psi_1(x), ..., \psi_m(x))'$ and

$Q_m(\theta_m) = \int f_0(x)\varphi_m(x)e^{\theta_m'\varphi_m(x)}dx$, then we have

$$\begin{aligned}
&\widehat{K}_m(\theta_m) \\
= \; & n^{-1}\sum_{i=1}^{n} \frac{c_m^2(\theta_m)\varphi_m(X_i)\varphi_m(X_i)' - c_m(\theta_m)\varphi_m(X_i)Q_m(\theta_m)'}{c_m^2(\theta_m)} \\
& -\frac{c_m(\theta_m)Q_m(\theta_m)\varphi_m(X_i)' - Q_m(\theta_m)Q_m(\theta_m)'}{c_m^2(\theta_m)},
\end{aligned}$$

and

$$\widehat{J}_m(\theta_m) = \frac{c_m(\theta_m)\int f_0\varphi_m\varphi_m' e^{\theta_m'\varphi_m}dx - Q_m(\theta_m)Q_m(\theta_m)'}{c_m^2(\theta_m)}.$$

Let $\hat{\theta}_m = (\hat{a}_1, ..., \hat{a}_m)'$ be the maximum likelihood estimator of $\theta_m = (a_1, ..., a_m)'$ maximizing $\sum_{i=1}^n \theta_m'\varphi_m(X_i) - n\log c_m(\theta_m)$. Then replacing θ_m in $\widehat{K}_m(\theta_m)$ and $\widehat{J}_m(\theta_m)$ by $\hat{\theta}_m$, we obtain d_m and hence DMA. The resultant density averaging estimator over the density family (A.51) is given by $\sum_{m=1}^M \hat{w}_m f_m(x, \hat{\theta}_m)$, where $\hat{w} = (\hat{w}_1, ..., \hat{w}_M)' = \text{argmin}_{w\in\mathcal{W}} C^{\text{DMA}}(w)$.

A.8 The explanations on the technical conditions

Let the random variable X have a bounded compact support $\mathcal{C} = [a, b]$ with a and b being two positive constants. Specially, we let the candidate model set $\mathcal{D} = \{f_1(x;\lambda) = \lambda e^{-\lambda x}, f_2(x;\mu,1) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2}\}\}$. Denote $\theta_0 = (\lambda_0, \mu_0)'$, where λ_0 and μ_0 are theoretically optimal parameters of λ and μ, respectively. Let $\aleph$ denote some neighbourhood of θ_0. In this special

case, it is clear that $f_1(x;\lambda)$ and $f_2(x;\mu,1)$ have lower and upper bounds for any $x \in \mathcal{C}$ and $\theta \in \aleph$. So, Conditions (C.1), (C.11) and (C.13) are satisfied.

Further, by some direct calculations, we obtain $\frac{\partial \log f_1(X;\lambda)}{\partial \lambda} = \frac{1}{\lambda} - X$, $\frac{\partial \log f_2(X;\mu,1)}{\partial \mu} = X - \mu$, $\frac{\partial \log f(X,\theta,w)}{\partial \lambda} = \frac{w_1(1-\lambda X)e^{-\lambda X}}{f(X,\theta,w)}$,

$$\frac{\partial^2 \log f(X,\theta,w)}{\partial \lambda^2} = \frac{-w_1 X e^{-\lambda X}(2-\lambda X)f(X,\theta,w) - [w_1(1-\lambda X)e^{-\lambda X}]^2}{f^2(X,\theta,w)},$$

$$\frac{\partial \log f(X,\theta,w)}{\partial \mu} = \frac{w_2 f_2(X;\mu,1)(X-\mu)}{f(X,\theta,w)},$$

$$\frac{\partial^2 \log f(X,\theta,w)}{\partial \mu^2} = \frac{w_2 f_2(X;\mu,1)[(X-\mu)^2-1]f(X,\theta,w) - [w_2 f_2(X;\mu,1)(X-\mu)]^2}{f^2(X,\theta,w)},$$

$\frac{\partial f_1(X;\lambda)}{\partial \lambda} = (1-\lambda X)e^{-\lambda X}$ and $\frac{\partial f_2(X;\mu,1)}{\partial \mu} = f_2(X;\mu,1)(X-\mu)$. Then, it is clear that $\sup_{x\in\mathcal{C},\theta\in\aleph}\left|\frac{\partial \log f_1(x;\lambda)}{\partial \lambda}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph}\left|\frac{\partial \log f_2(x;\mu,1)}{\partial \mu}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph,w\in\mathcal{W}}\left|\frac{\partial \log f(x,\theta,w)}{\partial \lambda}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph,w\in\mathcal{W}}\left|\frac{\partial^2 \log f(x,\theta,w)}{\partial \lambda^2}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph,w\in\mathcal{W}}\left|\frac{\partial \log f(x,\theta,w)}{\partial \mu}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph,w\in\mathcal{W}}\left|\frac{\partial^2 \log f(x,\theta,w)}{\partial \mu^2}\right|$, $\sup_{x\in\mathcal{C},\theta\in\aleph}\left|\frac{\partial f_1(x;\lambda)}{\partial \lambda}\right|$ and $\sup_{x\in\mathcal{C},\theta\in\aleph}\left|\frac{\partial f_2(x;\mu,1)}{\partial \mu}\right|$ are bounded. This indicates that Conditions (C.2), (C.4), (C.5) and (C.12) hold for $f_1(X;\lambda)$ and $f_2(X;\mu,1)$.

Since the trace of $\frac{\partial^2 \log f(X,\theta,w)}{\partial \theta \theta'}$ is $\frac{\partial^2 \log f(X,\theta,w)}{\partial \lambda^2} + \frac{\partial^2 \log f(X,\theta,w)}{\partial \mu^2}$ which is bounded for any $w \in \mathcal{W}$, Condition (C.3) is satisfied. We let $g(x) \neq f_1(x;\lambda)$, $g(x) \neq f_2(x;\mu,1)$ and $g(x)$ be not a mixture density function, then for any $w \in \mathcal{W}$, $g(x)$ often cannot be equal to $f(x,\theta,w)$. This implies that ξ_n is bounded away from zero and thus Condition (C.6) is satisfied.

Because p_{max} and M are fixed, Conditions (C.7), (C.8) and (C.10) are the same as Conditions (C.2), (C.4) and (C.6), respectively. Since $\hat{\theta}_m$ is

the maximum likelihood estimator of θ_m, Condition (C.9) can be rigorously proved under some regularity conditions (see, for example, White (1982)).

Also, the trace of $\frac{\partial F(X,\theta)}{\partial \theta'}$ is $\frac{\partial f_1(X;\lambda)}{\partial \lambda}+\frac{\partial f_2(X;\mu,1)}{\partial \mu}$ which is bounded, so Condition (C.14) is satisfied. Since p_{max} and M are fixed, Condition (C.15) is clearly satisfied. When n is large enough, $\sum_{i=1}^{n} F(X_i,\widehat{\theta})F(X_i,\widehat{\theta})'/n$ is often a positive definite matrix with $F(X,\theta)=(f_1(X;\lambda),f_2(X;\mu,1))'$, and then Condition (C.16) can be satisfied. Thus, in this special case, Conditions (C.1)-(C.16), which are imposed to establish the asymptotic optimality of our DMA estimator and the weight consistency, are satisfied.

In addition, since X_i, $i=1,...,n$, are independent and identically distributed, $\hat{\theta}^{[-i]}$ and $\hat{w}^{[-i]}$ often tend to θ_0 and w^0, respectively. This indicates that $f(X_i,\hat{\theta}^{[-i]},\hat{w}^{[-i]})$ is close to $f(x,\theta_0,w^0)$. Similarly, $f_h^{[-i]}(X_i)$ often tends to $Ef_h(x)$. So, $\tilde{f}^{[-i]}(X_i,\rho)$ is close to $\tilde{f}^0(X_i,\rho)$ which implies that Condition (C.17) can be satisfied. If we suppose that h tends to zero as $n\to\infty$, then we have $Ef_h(x)=g(x)+o(1)$ with $g(x)$ being bounded for any $x\in\mathcal{C}$ (see, e.g., Hansen, 2022). From Condition (C.1), we obtain that $f(x,\theta_0,w^0)$ has lower and upper bounds for any $x\in\mathcal{C}$. So, Condition (C.18) can be satisfied. When $p_{\max}$ and M are fixed and h tends to zero, $E\left[\log\tilde{f}^0(X,\rho)\right]$ often has a limit as $n\to\infty$. So, Condition (C.19) can be satisfied.

A.9 Discussion on the different density aggregation methods

The aggregation estimation for density function has been studied in related works such as Rigollet and Tsybakov (2007) (aggregation for squared loss), Dalalyan and Tsybakov (2012) (aggregation with infinite elements), and Dalalyan and Sebbar (2018) (aggregation for KL loss), among others. See also Tsybakov (2003) for aggregation estimation in the regression context. Especially, Dalalyan and Sebbar (2018) developed the optimal Kullback-Leibler aggregation in mixture density estimation and derived sharp oracle inequalities on risk bounds for the mixing density estimator, where the weights are estimated by maximizing the logarithmic likelihood function.

Both our paper and Dalalyan and Sebbar (2018) attempt to estimate the true density by averaging the candidate density functions in terms of KL distance. However, the work of Dalalyan and Sebbar (2018), as well as many other related researches in this topic (e.g., Dalalyan and Tsybakov (2012) and Butucea et al., 2017), was developed in the idealized context that the candidate densities are known, i.e., the pure aggregation framework (see also Rigollet and Tsybakov, 2007). Instead, our work considers the more practical situation where the candidate densities are unknown (because they contain unknown parameters) and hence need to be estimated based on the whole sample. This involves the development of aggregation

strategy because both the biases and variances of candidate density estimators should be considered. In such a case, we develop a novel data-driven weight choice criterion, where the criterion takes both the bias and variance into account and is an approximation to the KL distance between the true density and the averaging density estimator. Moreover, we have established its asymptotic optimality in the sense that the resultant averaging density estimator can achieve the minimum KL distance.

In fact, when we apply the method proposed in Dalalyan and Sebbar (2018) to the weight selection in our framework, the corresponding weight estimator is derived by maximizing the logarithmic score (LS) (Hall and Mitchell, 2007). The numerical results show that our proposed DMA procedure often performs better than LS; see our numerical results in Section 4 of the main text.

A.10 Figures

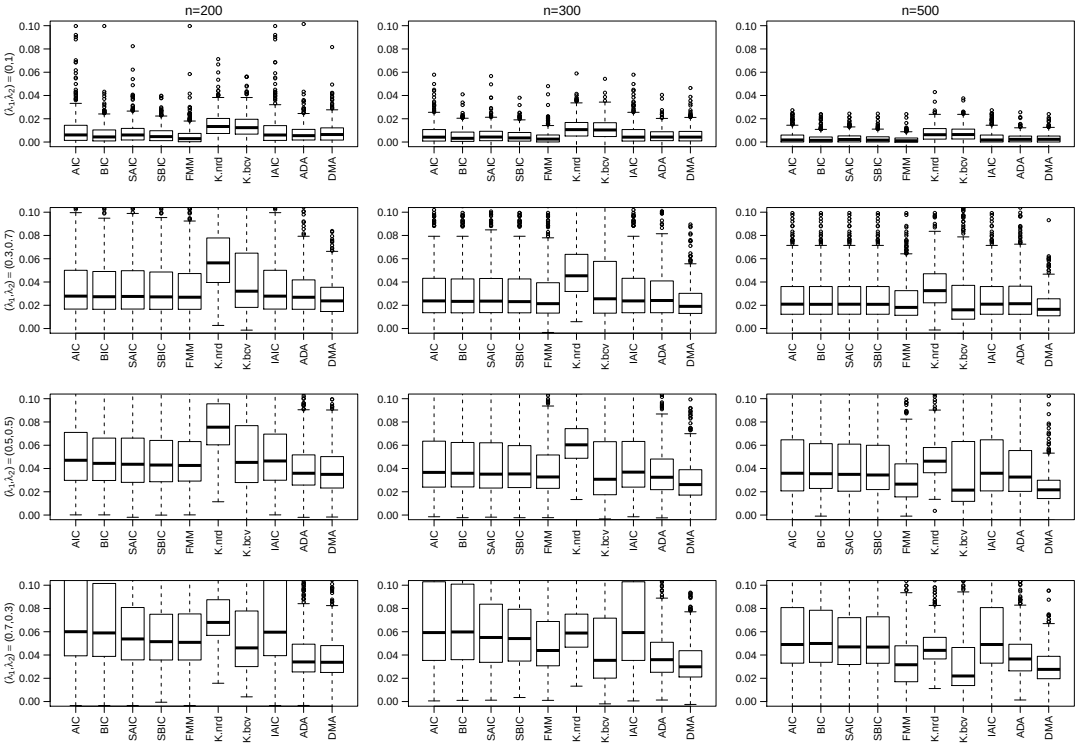

Figure 1: KL losses for various methods for Case 1 where the true distribution function of X is $\lambda_1 \times \mathrm{LN}\,(0.5, 0.5^2) + \lambda_2 \times \mathrm{N}\,(4, 0.5^2)$.

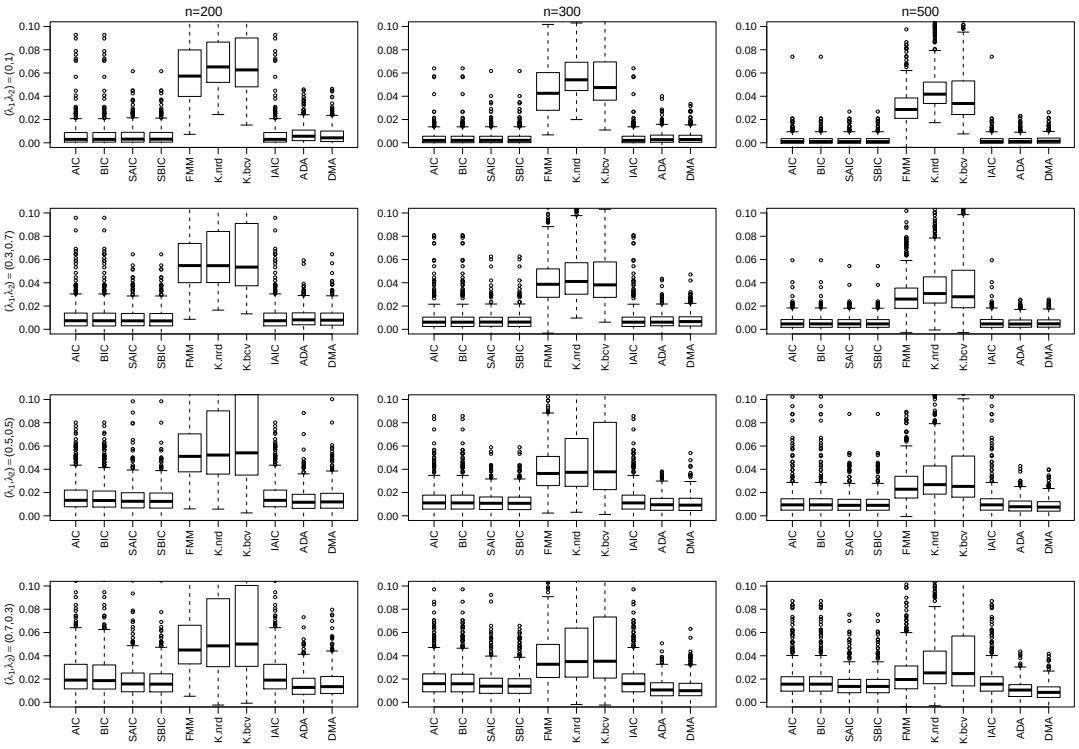

Figure 2: KL losses for various methods for Case 2 where the true distribution function of X is $\lambda_1 \times \text{LN}\,(0.5, 0.5^2) + \lambda_2 \times \text{Gamma}\,(2, 1)$.

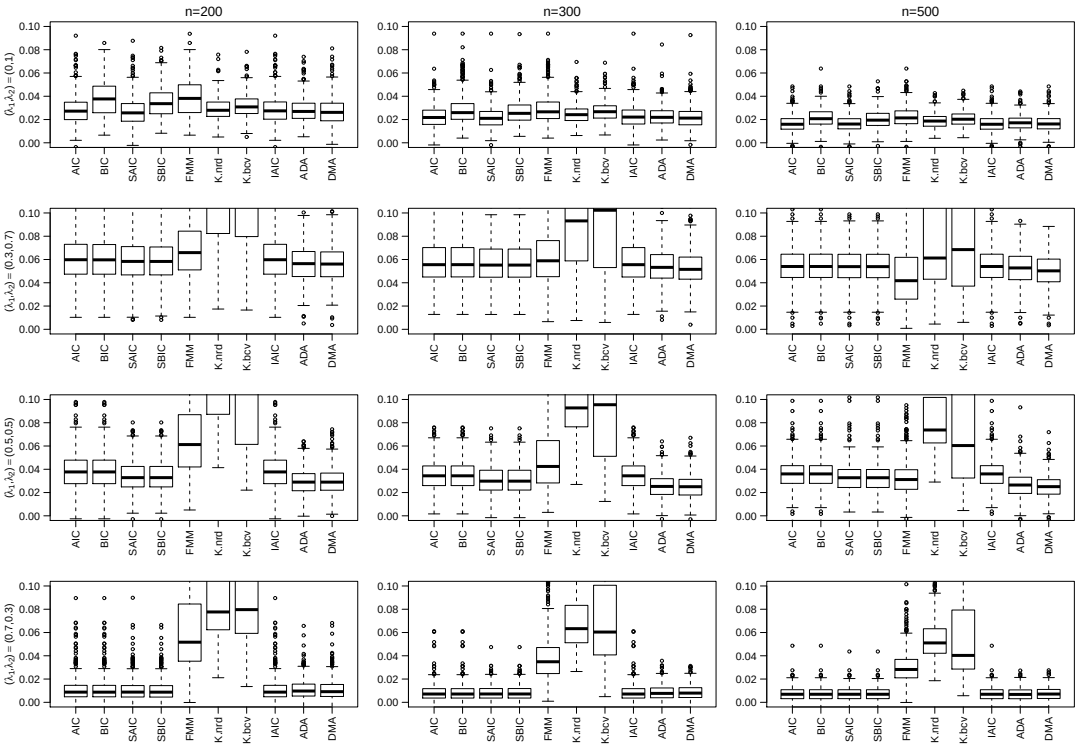

Figure 3: KL losses for various methods for Case 3 where the true distribution function of X is $\lambda_1 \times \text{LN}\,(0.5, 0.5^2) + \lambda_2 \times \text{Beta}\,(2, 2)$.

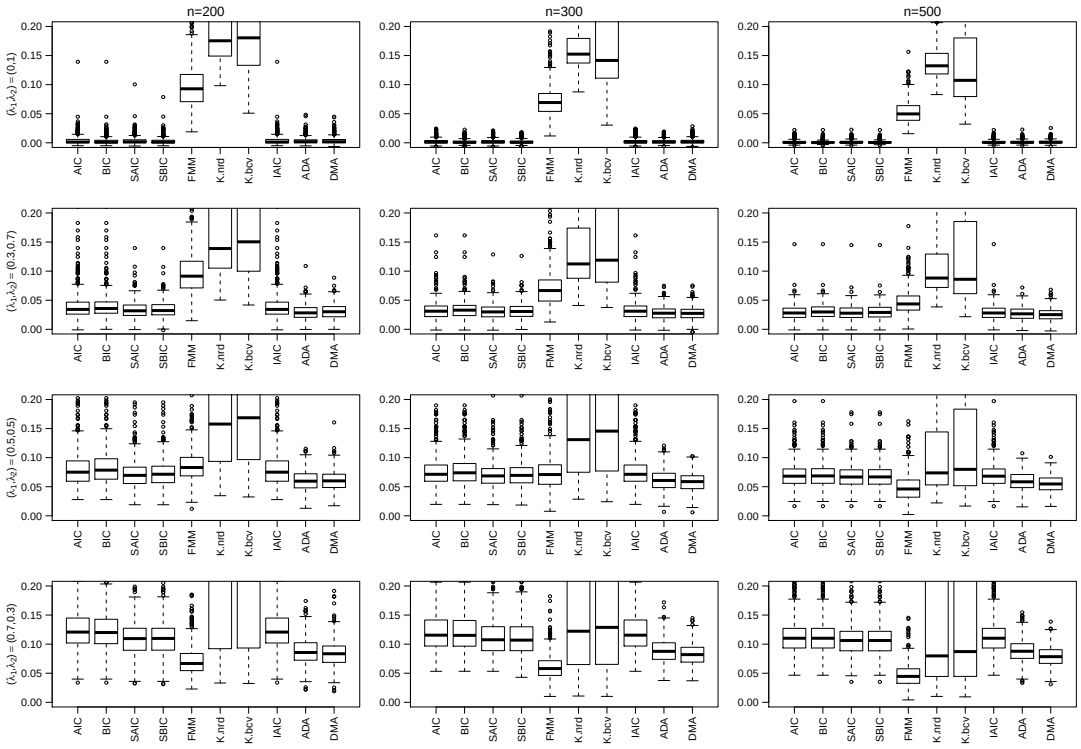

Figure 4: KL losses for various methods for Case 4 where the true distribution function of X is $\lambda_1 \times$Beta $(2,2) + \lambda_2 \times$ E(1).

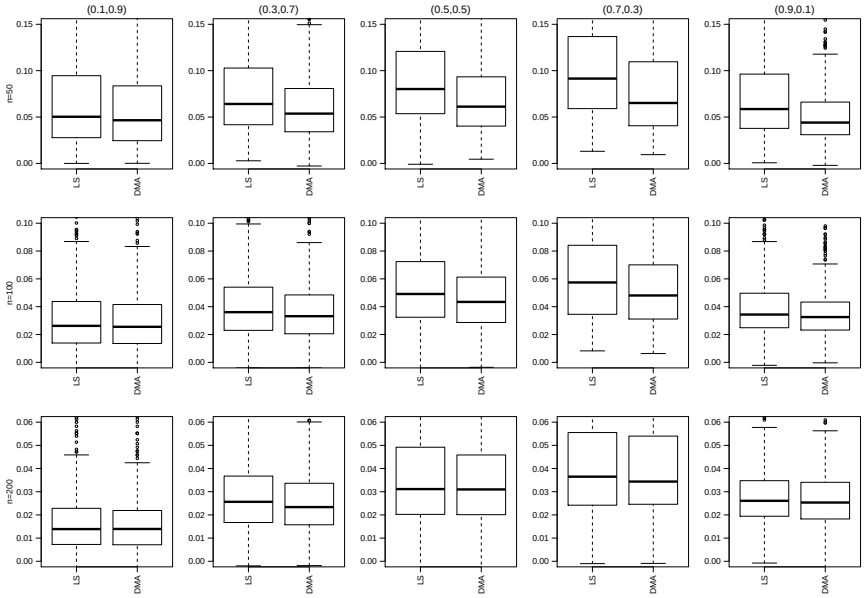

Figure 5: KL losses for DMA and the logarithmic scoring rule (LS) with $(\lambda_1, \lambda_2) \in \{(0.1, 0.9), (0.3, 0.7), (0.5, 0.5), (0.7, 0.3), (0.9, 0.1)\}$.

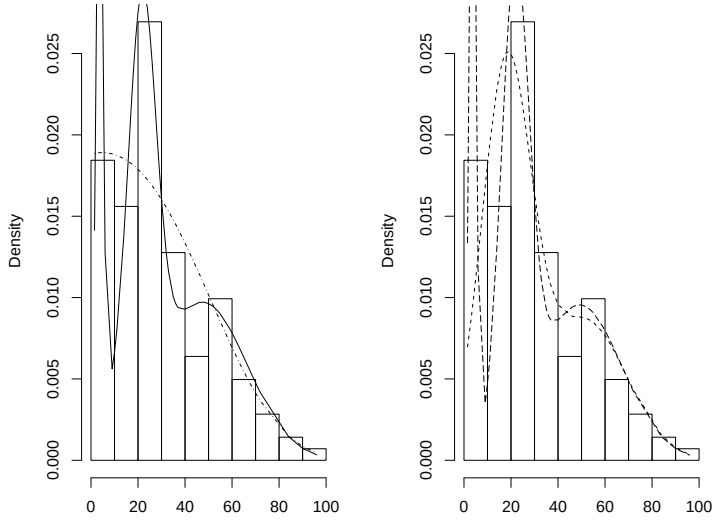

Figure 6: Histogram of life length data in ancient Egypt, and the Gompertz Model 1 (dotted dash line), DMA (solid line), and the mixture models with two and three mixing components (dash and long dash lines, respectively) based density estimations. The horizontal axis is the age at death.

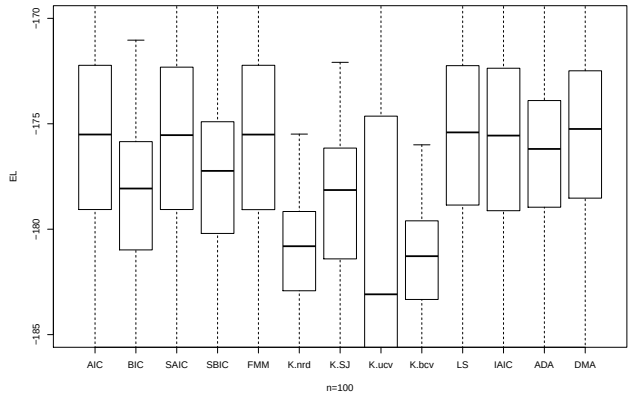

Figure 7: Boxplots for EL of each method based on 2000 replications for life length data.

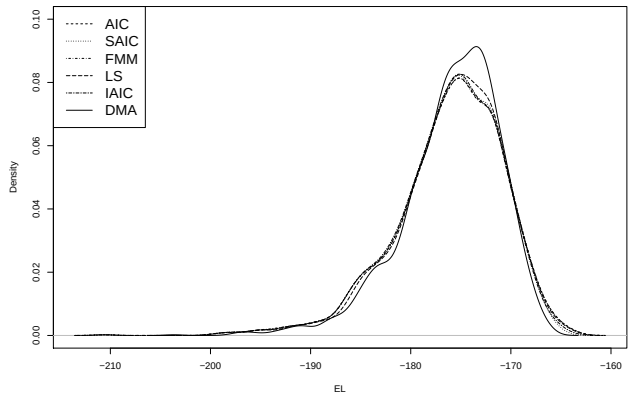

Figure 8: Comparison of DMA, LS, AIC, SAIC, IAIC & FMM in densities of ELs based on 2000 replications for life length data.

References

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press, Cambridge.

Butucea, C., Delmas, J.-F., Dutfoy, A., and Fischer, R. (2017). Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electronic Journal of Statistics*, 11:2258–2294.

Chen, J., Li, D., Linton, O., and Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113:919–932.

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

Dalalyan, A. and Sebbar, M. (2018). Optimal Kullback-Leibler aggregation in mixture density estimation by maximum likelihood. *Mathematical Statistics and Learning*, 1:1–35.

Dalalyan, A. S. and Tsybakov, A. B. (2012). Mirror averaging with sparsity priors. *Bernoulli*, 18:914–944.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman and Hall, London.

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23:1–13.

Hansen, B. E. (2022). *Probability and Statistics for Economists*. Princeton University Press, Princeton.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Unpublished manuscript.

Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16:260–280.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, pages 303–313. Springer Berlin Heidelberg.

Wan, A. T. K., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.