OPTIMAL MODEL AVERAGING BASED ON GENERALIZED METHOD OF MOMENTS

Supplementary Material

Xinyu Zhang

The following two lemmas will be used in the proofs of Proposition 1 and Theorem 1, respectively.

Lemma 1 *(Stein, 1981) Let $a \sim Normal(0,1)$ and $g(a) : \mathcal{R} \to \mathcal{R}$ be an indefinite integral of the Lebesgue measurable function $\dot{g}(a)$. Thus, $\dot{g}(a)$ is the derivative of $g(a)$. Suppose that $E|\dot{g}(a)| < \infty$. Then we have $E\{\dot{g}(a)\} = E\{ag(a)\}$.*

Lemma 2 *(Zhang, 2010; Gao et al., 2019) Let*

$$\widetilde{\mathbf{w}} = argmin_{\mathbf{w}\in\mathcal{W}} \left\{ L(w) + a_n(w) + b_n \right\},$$

where $a_n(\mathbf{w})$ is a term related to $\mathbf{w}$ and b_n is a term unrelated to $\mathbf{w}$. If

$$\sup_{\mathbf{w}\in\mathcal{W}} |a_n(\mathbf{w})|/L^*(\mathbf{w}) = o_p(1), \qquad \sup_{\mathbf{w}\in\mathcal{W}} |L(\mathbf{w}) - L^*(\mathbf{w})|/L^*(\mathbf{w}) = o_p(1),$$

and there exists a positive constant c and a positive integer N such that when

$n \geq N$, $\inf_{\mathbf{w}\in\mathcal{W}} L^*(\mathbf{w}) \geq c > 0$ *almost surely, then* $L(\widetilde{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w}) \to 1$ *in probability.*

S.1 Proof of Proposition 1

Let $f(\cdot)$ be a function with $f[\sqrt{n}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}] = \sqrt{n}\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \sqrt{n}\widehat{\boldsymbol{\mu}}$. It is seen that

$$
\begin{aligned}
& R(\mathbf{w}) \qquad\qquad (\text{S.1})\\
= \;& E\left([\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)]\right)\\
= \;& E\left([\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)]\right)\\
= \;& E\left([\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]\right) + E\left[\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}^{\mathrm{T}}\boldsymbol{\Omega}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}\right]\\
& +2E\left([\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}\right) \qquad\qquad (\text{S.2})
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left([\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}\right)\\
= \;& n^{-1}E\left([\sqrt{n}\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \sqrt{n}\widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}\sqrt{n}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}\right)\\
= \;& n^{-1}E\left(f[\sqrt{n}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}]\boldsymbol{\Omega}\sqrt{n}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\}\right)\\
= \;& n^{-1}\left[E\left\{f(\boldsymbol{\pi})^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\pi}\right\} + o(1)\right]\\
= \;& n^{-1}\left[E\left(\operatorname{trace}\left\{\frac{\partial f(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}^{\mathrm{T}}}\boldsymbol{\Omega}\mathbf{V}\right\}\right) + o(1)\right]\\
= \;& n^{-1}\left[E\left(\operatorname{trace}\left[\frac{\partial(\sqrt{n}\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \sqrt{n}\widehat{\boldsymbol{\mu}})}{\partial\sqrt{n}\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right\}^{\mathrm{T}}}\boldsymbol{\Omega}\boldsymbol{V}\right]\right) + o(1)\right]
\end{aligned}
$$

$$= \quad n^{-1}E\left(\text{trace}\left[\frac{\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathrm{w})\}}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\boldsymbol{\Omega}\boldsymbol{V}\right]\right) - n^{-1}\text{trace}(\boldsymbol{\Omega}\boldsymbol{V}) + o(n^{-1})$$

$$= \quad n^{-1}E\left(\text{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathrm{w})\}}{\partial\widehat{\boldsymbol{\theta}}(\mathrm{w})^{\mathrm{T}}}\boldsymbol{\Pi}_m^{\mathrm{T}}\frac{\partial\widehat{\boldsymbol{\theta}}_m}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\boldsymbol{\Omega}\boldsymbol{V}\right]\right) - n^{-1}\text{trace}(\boldsymbol{\Omega}\boldsymbol{V}) + o(n^{-1}),$$

where the third, fourth and fifth steps are from Lemma 1 and Conditions (C.1)-(C.2). The above two formulas imply (3.6). This completes the proof.

S.2 Proof of Proposition 2

It is implied by (2.5) that

$$\frac{\partial\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\}^{\mathrm{T}}\boldsymbol{\Omega}\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\}}{\partial\widehat{\boldsymbol{\theta}}_m} = \mathbf{0}, \tag{S.3}$$

which is

$$\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\right\} = \mathbf{0}. \tag{S.4}$$

Taking derivative of the both sides of (S.4) with respect to $\widehat{\boldsymbol{\mu}}^{\mathrm{T}}$, we have

$$\sum_{\tau=1}^{d_m}\boldsymbol{A}_\tau(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\left\{\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)\right\}\frac{\partial\widehat{\theta}_{m,\tau}}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}} + \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega} \tag{S.5}$$

$$-\sum_{\tau=1}^{d_m}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\frac{\partial\boldsymbol{\mu}(\boldsymbol{\Pi}_m^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_m)}{\partial\widehat{\theta}_{m,\tau}}\frac{\partial\widehat{\theta}_{m,\tau}}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}} = \mathbf{0}. \tag{S.6}$$

From the definitions of $\boldsymbol{D}_m$ and $\boldsymbol{B}_m$ in (3.11) and (3.12), Equation (S.5) is simplified to

$$\boldsymbol{D}_m\frac{\partial\widehat{\boldsymbol{\theta}}_m}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}} + \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega} - \boldsymbol{B}_m\frac{\partial\widehat{\boldsymbol{\theta}}_m}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}} = \mathbf{0},$$

which implies

$$(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} = -(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega},$$

which, along with the condition that $(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)$ is invertible, implies (3.13). This completes the proof.

S.3 Proofs of (3.16), (3.17), (3.20) and (3.21)

Let $\widehat{\mathbf{B}}_m = \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\boldsymbol{A}^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_m)$. Then, we have

$$\begin{aligned}
& \operatorname{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}}\boldsymbol{\Pi}_m^{\mathrm{T}}\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\boldsymbol{\Omega}\widehat{\boldsymbol{V}}\right] \\
= \ & \operatorname{trace}\left\{\sum_{m=1}^{M} w_m \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}\boldsymbol{\Pi}_m^{\mathrm{T}}(\widehat{\mathbf{B}}_m^{\mathrm{T}}\widehat{\mathbf{B}}_m)^{-1}\widehat{\mathbf{B}}_m^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\boldsymbol{\Omega}\widehat{\mathbf{V}}\right\} \\
= \ & \widehat{\sigma}^2\operatorname{trace}\left\{\sum_{m=1}^{M} w_m \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)^{\mathrm{T}}(\widehat{\mathbf{B}}_m^{\mathrm{T}}\widehat{\mathbf{B}}_m)^{-1}\widehat{\mathbf{B}}_m^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\right\} \\
= \ & \widehat{\sigma}^2\operatorname{trace}\left\{\sum_{m=1}^{M} w_m \boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)^{\mathrm{T}}(\widehat{\mathbf{B}}_m^{\mathrm{T}}\widehat{\mathbf{B}}_m)^{-1}\widehat{\mathbf{B}}_m^{\mathrm{T}}\right\} \\
= \ & \widehat{\sigma}^2\operatorname{trace}\left\{\sum_{m=1}^{M} w_m \widehat{\mathbf{B}}_m(\widehat{\mathbf{B}}_m^{\mathrm{T}}\widehat{\mathbf{B}}_m)^{-1}\widehat{\mathbf{B}}_m^{\mathrm{T}}\right\} \\
= \ & \widehat{\sigma}^2\sum_{m=1}^{M} w_m d_m, \qquad \text{(S.7)}
\end{aligned}$$

where the first step is from (3.13)-(3.15) and the second step is from (3.14)-(3.15). Hence, (3.16) is proved.

From (3.14) and (3.16), we have

$C(\mathbf{w})$

$$
\begin{aligned}
&= [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}] \\
&\quad +2n^{-1}\mathrm{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \boldsymbol{\Omega} \widehat{\boldsymbol{V}}\right] \\
&= [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}] + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{X}^{\mathrm{T}}\mathbf{y}\}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{X}^{\mathrm{T}}\mathbf{y}\} + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\left\{\boldsymbol{\theta}(\mathbf{w})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) + \mathbf{y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w})\right\} + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\|\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{y}\|^2 + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m - \mathbf{y}^{\mathrm{T}}\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right\}\mathbf{y},
\end{aligned}
$$

which is (3.17).

The proof of (3.20) is exactly the same as that of (3.16). For (3.21),

$$
\begin{aligned}
& C(\mathbf{w}) \\
&= [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}] \\
&\quad +2n^{-1}\mathrm{trace}\left[\sum_{m=1}^{M} w_m \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \boldsymbol{\Omega} \widehat{\boldsymbol{V}}\right] \\
&= [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \widehat{\boldsymbol{\mu}}] + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\{\mathbf{Z}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{Z}^{\mathrm{T}}\mathbf{y}\}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\{\mathbf{Z}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{Z}^{\mathrm{T}}\mathbf{y}\} + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\left\{\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) + \mathbf{y}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w})\right\} + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m \\
&= n^{-1}\|\mathbf{P}_{\mathbf{Z}}\mathbf{X}\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \mathbf{y}\|^2 + 2n^{-1}\widehat{\sigma}^2 \sum_{m=1}^{M} w_m d_m - \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\mathbf{y}.
\end{aligned}
$$

Hence, (3.21) is proved.

S.4 Examples where Conditions (C.3)-(C.5) and (C.7) are satisfied

We first consider the example with the linear regression candidate models, which are described in Remark 1 detailedly. In this example, $\mathbf{V} = \sigma^2 E(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}})$, $\widehat{\mathbf{V}} = \widehat{\sigma}^2\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$, $\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}/\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}\ |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$, and

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}_m &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{y}\\ &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\epsilon})\\ &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} + (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}.\end{aligned}$$

Therefore, when $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ converges to a positive definite matrix, $\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}/n = o_p(1)$ and $\widehat{\sigma}^2 - \sigma^2 = o_p(1)$, Conditions (C.3)-(C.5) and (C.7) are satisfied in this example.

Second, we consider the example with linear regression models with instrumental variables, which are described in Remark 2 detailedly. In this example, $\boldsymbol{\Omega} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n)^{-1}$, $\mathbf{V} = \sigma^2 E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})$ with $\mathbf{Z}_i^{\mathrm{T}}$ being the i^{th} row of $\mathbf{Z}$, $\widehat{\mathbf{V}} = \widehat{\sigma}^2\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$, $\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}/\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}\ |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}} = \mathbf{Z}^{\mathrm{T}}\mathbf{X}/n$, and

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}_m &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{y}\\ &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}(\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\epsilon})\\ &= (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\boldsymbol{\theta} + (\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}})^{-1}\boldsymbol{\Pi}_m\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\epsilon}.\end{aligned}$$

Therefore, when $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ converges to a positive definite matrices, $\mathbf{Z}^{\mathrm{T}}\mathbf{X}/n$

converges to a matrix with full column rank, $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\epsilon}/n = o_p(1)$ and $\widehat{\sigma}^2 - \sigma^2 = o_p(1)$, Conditions (C.3)-(C.5) and (C.7) are satisfied in this example.

S.5 Proof of Theorem 1

It is well-known that the following equalities are satisfied for any matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ with identical dimensions (see, for example, Li (1987)):

$$\lambda_{\max}(\mathbf{B}_1 + \mathbf{B}_2) \le \lambda_{\max}(\mathbf{B}_1) + \lambda_{\max}(\mathbf{B}_2) \text{ and } \lambda_{\max}(\mathbf{B}_1\mathbf{B}_2) \le \lambda_{\max}(\mathbf{B}_1)\lambda_{\max}(\mathbf{B}_2), \tag{S.8}$$

where the definition of $\lambda_{\max}(\cdot)$ is in Condition (C.5).

Now, we prove that uniformly for any $m \in \{1, \ldots, M\}$,

$$\lambda_{\max}\left(\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\right) = O_p(1). \tag{S.9}$$

Let $\mathbf{P}_{\mathbf{BD}} = (\boldsymbol{D}_m - \boldsymbol{B}_m)\left\{(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\right\}^{-1}(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}$. By (3.13), (S.8), the assumption that $(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)$ is invertible, and the truth that $\boldsymbol{\Omega}$ is a positive definite matrix, we have that uniformly for $m \in \{1, \ldots, M\}$,

$$\begin{aligned}
& \lambda_{\max}\left(\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\right) \\
= \;& \lambda_{\max}^{1/2}\left(\frac{\partial \widehat{\boldsymbol{\theta}}_m^{\mathrm{T}}}{\partial \widehat{\boldsymbol{\mu}}}\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\right) \\
= \;& \lambda_{\max}^{1/2}\left(\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\left\{(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\right\}^{-2}(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\right) \\
\le \;& \lambda_{\max}^{1/2}\left(\left\{(\boldsymbol{D}_m - \boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m - \boldsymbol{B}_m)\right\}^{-1}\right)\lambda_{\max}^{1/2}\left(\boldsymbol{\Omega}^{\mathrm{T}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)^{\mathrm{T}}\mathbf{P}_{\mathbf{BD}}\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\boldsymbol{\Omega}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \lambda_{\max}^{1/2}\left(\left\{(\boldsymbol{D}_m-\boldsymbol{B}_m)^{\mathrm{T}}(\boldsymbol{D}_m-\boldsymbol{B}_m)\right\}^{-1}\right)\lambda_{\max}^{1/2}\left(\mathbf{P}_{\mathbf{BD}}\right)\lambda_{\max}\left(\boldsymbol{A}(\widehat{\boldsymbol{\theta}}_m)\right)\lambda_{\max}(\boldsymbol{\Omega}) \\
&= O(1), \qquad \text{(S.10)}
\end{aligned}
$$

hence, (S.9) is proved.

Let

$$
\boldsymbol{H}_m = 2n^{-1}\frac{\partial\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial\widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}}\boldsymbol{\Pi}_m^{\mathrm{T}}\frac{\partial\widehat{\boldsymbol{\theta}}_m}{\partial\widehat{\boldsymbol{\mu}}^{\mathrm{T}}}\boldsymbol{\Omega}\widehat{\boldsymbol{V}} \quad \text{and} \quad \boldsymbol{H}(\mathbf{w}) = \sum_{m=1}^{M} w_m\boldsymbol{H}_m.
$$

It is seen that

$$
\begin{aligned}
& C(\mathbf{w}) \\
=\;& [\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\widehat{\boldsymbol{\mu}}]^{\mathrm{T}}\boldsymbol{\Omega}[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\widehat{\boldsymbol{\mu}})] + \mathrm{trace}\{\boldsymbol{H}(\mathbf{w})\} \\
=\;& \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)+\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\right]^{\mathrm{T}}\boldsymbol{\Omega}\left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)+\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\right] \\
& +\mathrm{trace}\{\boldsymbol{H}(\mathbf{w})\} \\
=\;& L(\mathbf{w}) + 2\left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}}\boldsymbol{\Omega}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\} + \{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\}^{\mathrm{T}}\boldsymbol{\Omega}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\} \\
& +\mathrm{trace}\{\boldsymbol{H}(\mathbf{w})\} \\
=\;& L(\mathbf{w}) + 2\left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}-\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}}\boldsymbol{\Omega}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}) \\
& +2\left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}-\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}}\boldsymbol{\Omega}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\} \\
& +\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Omega}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\} + \mathrm{trace}\{\boldsymbol{H}(\mathbf{w})\}, \qquad \text{(S.11)}
\end{aligned}
$$

where the term $\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}})^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)-\widehat{\boldsymbol{\mu}}\}$ is unrelated to $\mathbf{w}$, and

$$
L(\mathbf{w})
$$

$$
\begin{aligned}
&= \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right] \\
&= \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} + \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega} \\
&\quad \times \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} + \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right] \\
&= L^*(\mathbf{w}) + \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right] \\
&\quad +2 \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right]. \qquad \text{(S.12)}
\end{aligned}
$$

In addition, from Condition (C.6), we know that there exists a positive constant c and a positive integer N such that when $n \geq N$, $\inf_{\mathbf{w}\in\mathcal{W}} L^*(\mathbf{w}) \geq c > 0$ almost surely. Hence, by Lemma 2, to prove (4.2) it is sufficient to verify that

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|L^*(\mathbf{w})^{-1} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]\right| = o_p(1), \qquad \text{(S.13)}$$

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|L^*(\mathbf{w})^{-1} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right]\right| = o_p(1), \qquad \text{(S.14)}$$

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|L^*(\mathbf{w})^{-1} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\}\right]^{\mathrm{T}} \boldsymbol{\Omega}\{\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\}\right| = o_p(1), \qquad \text{(S.15)}$$

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|L^*(\mathbf{w})^{-1} \left[\boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} - \boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega}\{\boldsymbol{\mu}_{\text{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\}\right| = o_p(1), \qquad \text{(S.16)}$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}} |n^{-1}L^*(\mathbf{w})^{-1}\text{trace}\{\boldsymbol{H}(\mathbf{w})\}| = o_p(1). \qquad \text{(S.17)}$$

By Taylor's expansion, we obtain that

$$
\begin{aligned}
& \left\| \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}\{\boldsymbol{\theta}^*(\mathbf{w})\} \right\|^2 \\
= & \left\| \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}} \{\widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\} \right\|^2 \\
\leq & \lambda_{\max} \left[\frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}^*_{\mathbf{w}}} \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}^{\mathrm{T}}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})} |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}^*_{\mathbf{w}}} \right] \left\| \widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w}) \right\|^2 \\
\leq & \lambda^2_{\max} \left[\frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} |_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}^*_{\mathbf{w}}} \right] \left\| \widehat{\boldsymbol{\theta}}(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w}) \right\|^2 \\
= & O_p(n^{-1}Mp), \qquad \text{(S.18)}
\end{aligned}
$$

where $\widetilde{\boldsymbol{\theta}}^*_{\mathbf{w}}$ is a vector between $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ and $\boldsymbol{\theta}^*(\mathbf{w})$ and can be related to $\mathbf{w}$, the third step is from (S.8), and the last step is from Conditions (C.4) and (C.5).

From (S.18) and Condition (C.6), we can obtain (S.13)-(S.14). From (S.18) and Conditions (C.1), (C.3) and (C.6), we can obtain (S.15). From Conditions (C.1), (C.3) and (C.6), we can obtain (S.16).

It is seen that

$$
\begin{aligned}
& \operatorname{trace}\{\boldsymbol{H}(\mathbf{w})\} \\
\leq & \max_{1\leq m\leq M} \operatorname{trace}(\boldsymbol{H}_m) \\
= & 2^{-1} \max_{1\leq m\leq M} \operatorname{trace}(\boldsymbol{H}_m + \boldsymbol{H}_m^{\mathrm{T}}) \\
\leq & 2^{-1} \max_{1\leq m\leq M} \operatorname{rank}(\boldsymbol{H}_m + \boldsymbol{H}_m^{\mathrm{T}}) \lambda_{\max}(\boldsymbol{H}_m + \boldsymbol{H}_m^{\mathrm{T}}) \\
\leq & 2 \max_{1\leq m\leq M} \operatorname{rank}(\boldsymbol{H}_m) \lambda_{\max}(\boldsymbol{H}_m)
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2 \max_{1\leq m\leq M} \mathrm{rank}(\boldsymbol{H}_m) 2n^{-1} \max_{1\leq m\leq M} \lambda_{\max} \left[\frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \boldsymbol{\Pi}_m^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \boldsymbol{\Omega} \widehat{\boldsymbol{V}} \right] \\
&\leq 4n^{-1} p \max_{1\leq m\leq M} \lambda_{\max} \left[\frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \right] \lambda_{\max} \left(\boldsymbol{\Pi}_m^{\mathrm{T}} \right) \lambda_{\max} \left(\frac{\partial \widehat{\boldsymbol{\theta}}_m}{\partial \widehat{\boldsymbol{\mu}}^{\mathrm{T}}} \right) \\
&\quad \times \lambda_{\max} \left(\boldsymbol{\Omega} \right) \lambda_{\max} \left(\widehat{\boldsymbol{V}} \right) \\
&= O_p(p/n),
\end{aligned} \tag{S.19}
$$

where the fourth and sixth steps use (S.8) and the last step uses (S.9) and Conditions (C.3) and (C.5). Now, by (S.19) and Condition (C.6), we can obtain (S.17). As stated in above (S.13), the optimality (4.2) is implied by (S.13)-(S.17) This completes the proof.

S.6 Proof of Theorem 2

Let

$$
\boldsymbol{G}(\mathbf{w}) = \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}^{\mathrm{T}}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})} \Big|_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}^*} \boldsymbol{\Omega} \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})^{\mathrm{T}}} \Big|_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}^*}
$$

and

$$
\boldsymbol{g}(\mathbf{w}) = \frac{\partial \boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\}^{\mathrm{T}}}{\partial \widehat{\boldsymbol{\theta}}(\mathbf{w})} \Big|_{\widehat{\boldsymbol{\theta}}(\mathbf{w})=\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}^*} \boldsymbol{\Omega} \left\{ \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}} \right\},
$$

where $\widetilde{\boldsymbol{\theta}}_{\mathbf{w}}^*$ is defined following (S.18). It is seen that

$$
\begin{aligned}
&C(\mathbf{w}) \\
&= \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) + \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}} \right\}^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathbf{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) + \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}} \right\} \\
&\quad + \mathrm{trace}\{\boldsymbol{H}(\mathbf{w})\}
\end{aligned}
$$

$$
\begin{aligned}
&= \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathrm{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega} \left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathrm{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right] \\
&\quad +2\left[\boldsymbol{\mu}\{\widehat{\boldsymbol{\theta}}(\mathrm{w})\} - \boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0)\right]^{\mathrm{T}} \boldsymbol{\Omega}\left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\} \\
&\quad +\mathrm{trace}\{\boldsymbol{H}(\mathrm{w})\} + \left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\}^{\mathrm{T}} \boldsymbol{\Omega}\left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\} \\
&= \{\widehat{\boldsymbol{\theta}}(\mathrm{w}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{G}(\mathrm{w})\{\widehat{\boldsymbol{\theta}}(\mathrm{w}) - \boldsymbol{\theta}_0\} + 2\{\widehat{\boldsymbol{\theta}}(\mathrm{w}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{g}(\mathrm{w}) + \mathrm{trace}\{\boldsymbol{H}(\mathrm{w})\} \\
&\quad + \left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\}^{\mathrm{T}} \boldsymbol{\Omega}\left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\}, \qquad \text{(S.20)}
\end{aligned}
$$

where the first step is from the second step of (S.11) and the last step is from Taylor's expansion. Recall that $\mathbf{w}_{\widetilde{m}}$ is a weight vector in which the $\widetilde{m}^{th}$ component is one and the other are zeros. From (4.1), (S.19), Conditions (C.1) and (C.3), and the second step of (S.20), we have

$$
C(\mathbf{w}_{\widetilde{m}}) = \left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\}^{\mathrm{T}} \boldsymbol{\Omega}\left\{\boldsymbol{\mu}_{\mathrm{true}}(\boldsymbol{\theta}_0) - \widehat{\boldsymbol{\mu}}\right\} + O_p(n^{-1}p) = O_p(n^{-1}p). \qquad \text{(S.21)}
$$

From (S.19), Condition (C.1) and the third step of (S.20), we have

$$
C(\widehat{\mathbf{w}}) = \{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{G}(\widehat{\mathbf{w}})\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\} + 2\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{g}(\widehat{\mathbf{w}}) + O_p(n^{-1}p).
$$

Combining the above equations and $C(\widehat{\mathbf{w}}) \leq C(\mathbf{w}_{\widetilde{m}})$, we have

$$
\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{G}(\widehat{\mathbf{w}})\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\} + 2\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{g}(\widehat{\mathbf{w}}) + O_p(n^{-1}p) \leq O_p(n^{-1}p),
$$

from which and Condition (C.7), we further have

$$
\begin{aligned}
\kappa_2\|\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\|^2 &\leq -2\{\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\}^{\mathrm{T}}\boldsymbol{g}(\widehat{\mathbf{w}}) - O_p(n^{-1}p) + O_p(n^{-1}p) \\
&\leq 2\|\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\|\|\boldsymbol{g}(\widehat{\mathbf{w}})\| + O_p(n^{-1}p), \qquad \text{(S.22)}
\end{aligned}
$$

by which, we further have

$$\left\{\|\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_0\| - \kappa_2^{-1}\|\boldsymbol{g}(\widehat{\mathbf{w}})\|\right\}^2 \leq \kappa_2^{-2}\|\boldsymbol{g}(\widehat{\mathbf{w}})\|^2 + O_p(n^{-1}p). \qquad \text{(S.23)}$$

From Conditions (C.1), (C.3) and (C.5), it is easily to obtain $\|\boldsymbol{g}(\widehat{\mathbf{w}})\| = O_p(n^{-1/2}p^{1/2})$, which along with (S.23), implies (4.3). This completes the proof.

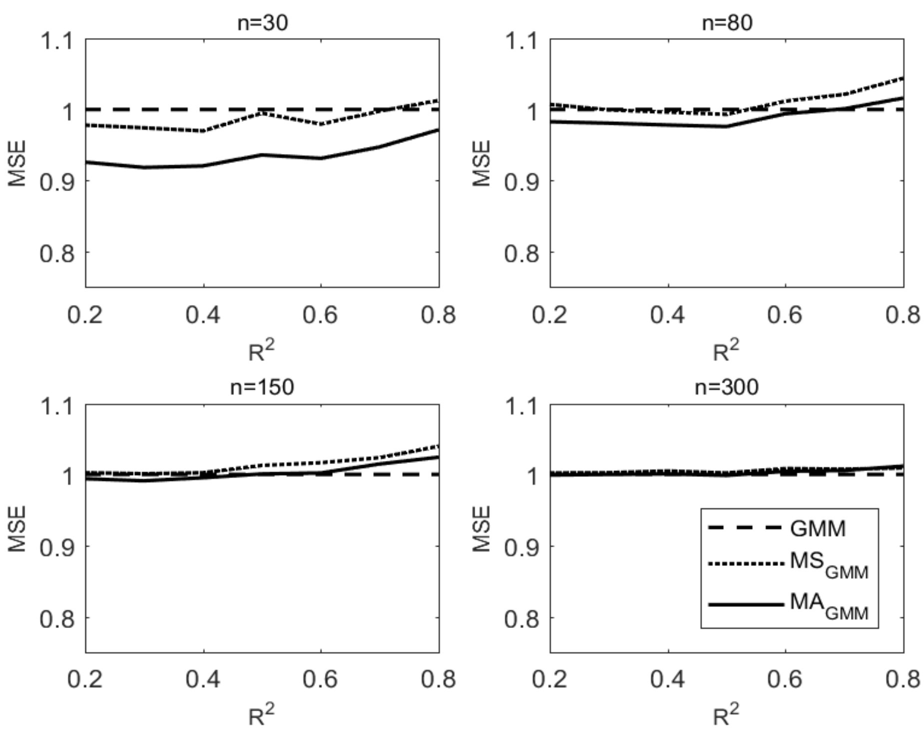

Figure S.1: MSE in simulation Design I, with $\widetilde{R}^2 = 0.5$.

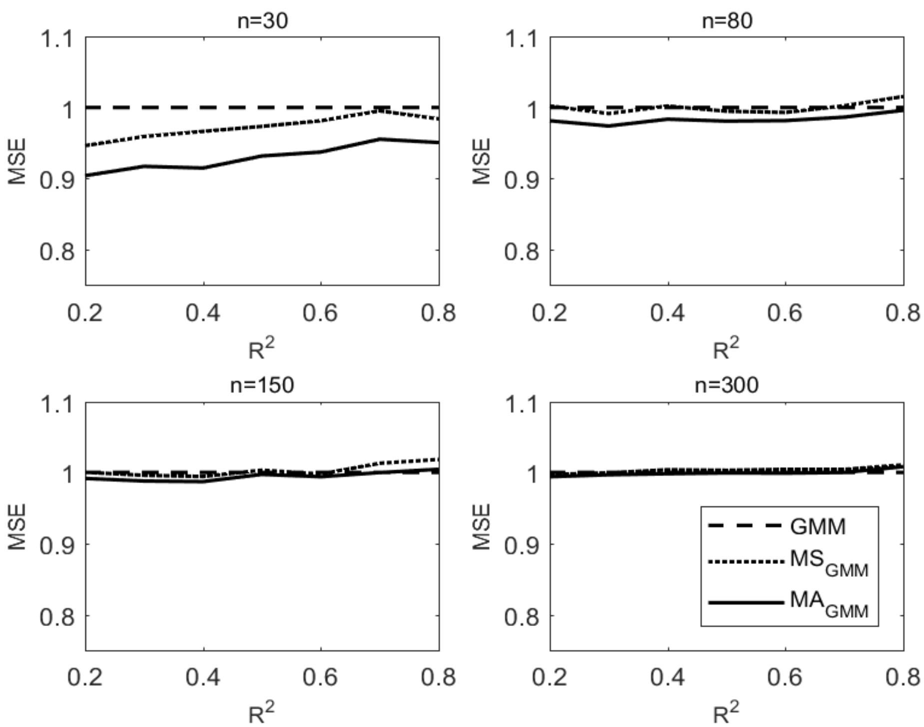

Figure S.2: MSE in simulation Design I, with $\widetilde{R}^2 = 0.8$.

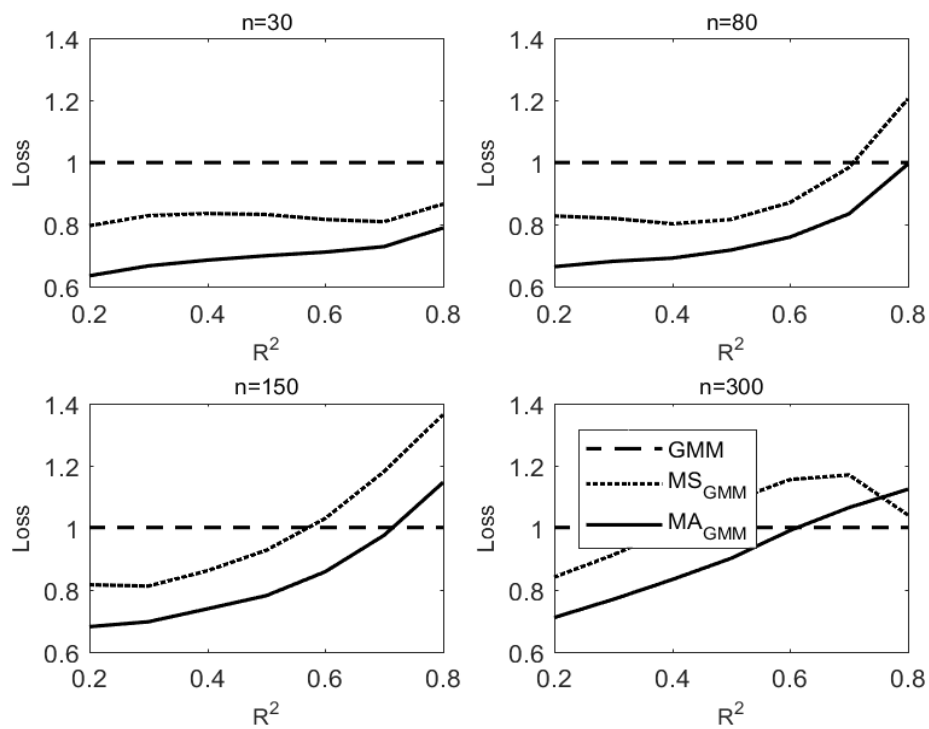

Figure S.3: Loss in simulation Design II, with $\widetilde{R}^2 = 0.5$.

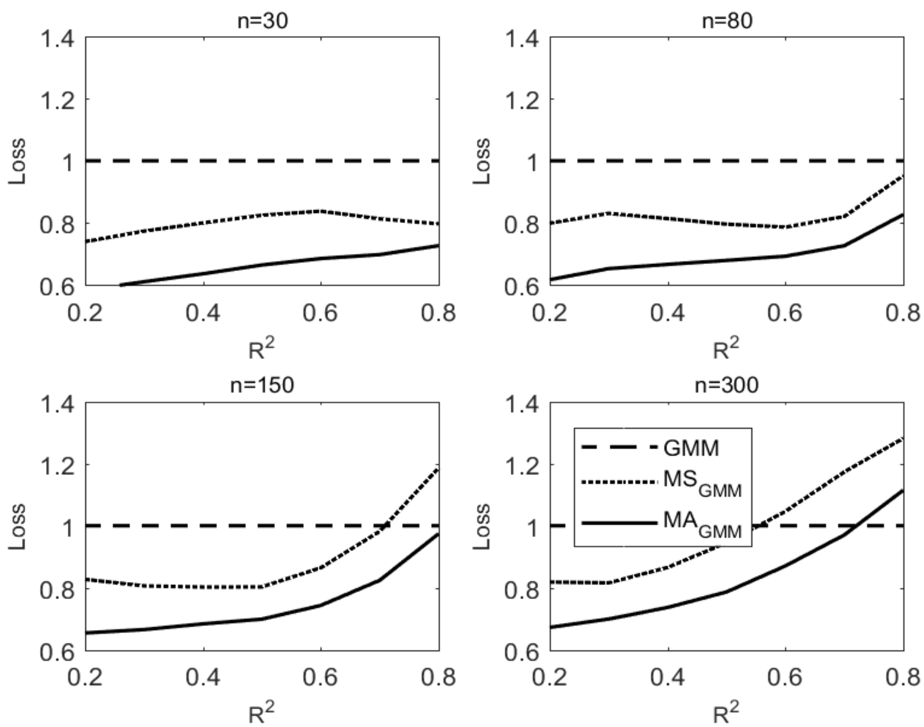

Figure S.4: Loss in simulation Design II, with $\widetilde{R}^2 = 0.8$.

References

GAO, Y., ZHANG, X., WANG, S., CHONG, T. T.-L. & ZOU, G. (2019). Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics* **71**, 275–306.

LI, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics* **15**, 958–975.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **153**, 1135–1151.

ZHANG, X. (2010). Model averaging and its applications. *Ph.D. Thesis* , Academy of Mathematics and Systems Science, Chinese Academy of Sciences.