Robust jackknife model averaging

Kang You[a,b], Miaomiao Wang[c] and Guohua Zou[a,¶]

[a] *Capital Normal University*
[b] *University of Kent*
[c] *Beijing University of Chinese Medicine*

Supplementary Material

The supplementary material contains the robustness property of the RJMA estimator, the proofs of theorems and additional simulation studies.

S1. Robustness property

In this section, we focus on exploring the robustness property of the RJMA estimator by means of influence function. Furthermore, an empirical prediction influence function is applied to evaluate the quantitative robustness of the proposed RJMA estimator.

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

S1.1 Influence function

It is well-known that influence function defined by Hampel (1968) is an important tool to measure the sensitivity of an estimator to small amounts of contamination in the distribution. Let $\Delta_{\boldsymbol{z}}$ be a Dirac measure putting all its mass at a fixed point $\boldsymbol{z} = (y^*, \boldsymbol{x}^{*T})^T \in \mathbb{R}^{p+1}$. Define $\breve{\Theta}_{(m)} = \arg\inf_{\Theta_{(m)}} \int h(\boldsymbol{x}_{(m)})\rho(y - \boldsymbol{x}_{(m)}^T\Theta_{(m)})d\mathrm{G}$ and $\breve{\Theta}_{\epsilon(m)} = \arg\inf_{\Theta_{(m)}} \int h(\boldsymbol{x}_{(m)})\rho(y - \boldsymbol{x}_{(m)}^T\Theta_{(m)})d\mathrm{G}_\epsilon$, where G is the joint distribution of $(y, \boldsymbol{x}^T)^T \in \mathbb{R}^{p+1}$ and $\mathrm{G}_\epsilon = (1-\epsilon)\mathrm{G} + \epsilon\Delta_{\boldsymbol{z}}$ is the mixture distribution of G and $\Delta_{\boldsymbol{z}}$ with $\epsilon \in (0, 1)$. By definition, the influence function of $\breve{\Theta}_{(m)}$ at the point $\boldsymbol{z} \in \mathbb{R}^{p+1}$ is $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)}) = \lim_{\epsilon \to 0^+} (\breve{\Theta}_{\epsilon(m)} - \breve{\Theta}_{(m)})/\epsilon$, provided that the limit exists for $m = 1, ..., M$. In other words, the influence function $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})$ describes the effect caused by infinitesimal contamination at the point $\boldsymbol{z}$ on $\breve{\Theta}_{(m)}$.

In the following, we define the influence function of the RJMA estimator. As we did in Section 3.2, we construct $\widetilde{\Theta}_{(m)}$ and $\widetilde{\Theta}_{\epsilon(m)}$ as two p-dimensional vectors which consist of all the elements in $\breve{\Theta}_{(m)}$ and $\breve{\Theta}_{\epsilon(m)}$, respectively, and $p - k_m$ zeroes. So it is obvious that $\sum_{m=1}^M w_m \boldsymbol{x}^T \widetilde{\Theta}_{(m)} = \sum_{m=1}^M w_m \boldsymbol{x}_{(m)}^T \breve{\Theta}_{(m)}$ and $\sum_{m=1}^M w_m \boldsymbol{x}^T \widetilde{\Theta}_{\epsilon(m)} = \sum_{m=1}^M w_m \boldsymbol{x}_{(m)}^T \breve{\Theta}_{\epsilon(m)}$. For $\widetilde{\Theta}_{(m)}$, the corresponding influence function $\mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)})$ is a p-dimensional vector which contains all the elements of $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})$ and $p - k_m$ zeroes. For any given weight vector $\boldsymbol{w} \in \mathcal{W}$, the model averaging parameter has the form

of $\sum_{m=1}^{M} w_m \widetilde{\Theta}_{(m)}$. So we define

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathcal{W}}{\arg\inf} \int h(\boldsymbol{x})\rho\left(y - \boldsymbol{x}^T \sum_{m=1}^{M} w_m \widetilde{\Theta}_{(m)}\right) d\mathrm{G}$$

and

$$\boldsymbol{w}_\epsilon^* = \underset{\boldsymbol{w}\in\mathcal{W}}{\arg\inf} \int h(\boldsymbol{x})\rho\left(y - \boldsymbol{x}^T \sum_{m=1}^{M} w_m \widetilde{\Theta}_{\epsilon(m)}\right) d\mathrm{G}_\epsilon.$$

Similar to the definition of $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})$, we define the influence function of the RJMA estimator at a point $\boldsymbol{z} \in \mathbb{R}^{p+1}$ as

$$\mathrm{IF}\left(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)}\right) = \lim_{\epsilon\to 0^+} \frac{\sum_{m=1}^{M} w_{\epsilon,m}^* \widetilde{\Theta}_{\epsilon(m)} - \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)}}{\epsilon}, \tag{S1.1}$$

where the limit is assumed to exist, and w_m^* and $w_{\epsilon,m}^*$ are the m^{th} elements of $\boldsymbol{w}^*$ and $\boldsymbol{w}_\epsilon^*$, respectively (see also Wang et al. (2024)).

To obtain the expression of the influence function (S1.1), we list some notations and regularity conditions. Denote $\boldsymbol{\pi}_\epsilon^* = (w_{\epsilon,1}^*, ..., w_{\epsilon,M-1}^*)^T$, $\boldsymbol{B}_\epsilon = \left(a_{\epsilon(1)}, ..., a_{\epsilon(M-1)}\right)^T$ with $a_{\epsilon(m)} = \boldsymbol{x}^T(\widetilde{\Theta}_{\epsilon(m)} - \widetilde{\Theta}_{\epsilon(M)})$, and $\boldsymbol{B}_\epsilon^* = (a_{\epsilon(1)}^*, ..., a_{\epsilon(M-1)}^*)^T$ with $a_{\epsilon(m)}^* = \boldsymbol{x}^{*T}(\widetilde{\Theta}_{\epsilon(m)} - \widetilde{\Theta}_{\epsilon(M)})$, then we have

$$\boldsymbol{x}^T \sum_{m=1}^{M-1} w_{\epsilon,m}^* (\widetilde{\Theta}_{\epsilon(m)} - \widetilde{\Theta}_{\epsilon(M)}) = \boldsymbol{\pi}_\epsilon^{*T} \boldsymbol{B}_\epsilon.$$

Define $\boldsymbol{H} = (h_{(1)}, ..., h_{(M-1)})^T$ with $h_{(m)} = \boldsymbol{x}^T\{\mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)}) - \mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(M)})\}$. Let $\boldsymbol{B}$, $\boldsymbol{B}^*$ and $\boldsymbol{\pi}^*$ be the counterparts of $\boldsymbol{B}_\epsilon$, $\boldsymbol{B}_\epsilon^*$ and $\boldsymbol{\pi}_\epsilon^*$ with $\widetilde{\Theta}_{(m)}$ and $\boldsymbol{w}^*$ in the places of $\widetilde{\Theta}_{\epsilon(m)}$ and $\boldsymbol{w}_\epsilon^*$, respectively. Denote $\boldsymbol{A}_{(m)} = \int h(\boldsymbol{x}_{(m)})\rho_2(y - \boldsymbol{x}_{(m)}^T \breve{\Theta}_{(m)})\boldsymbol{x}_{(m)}\boldsymbol{x}_{(m)}^T d\mathrm{G}$ and $\boldsymbol{D} = \int h(\boldsymbol{x})\rho_2(y - \boldsymbol{x}^T \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)})\boldsymbol{B}\boldsymbol{B}^T d\mathrm{G}$.

Assumption S1. *ψ has a first-order derivative ρ_2 almost surely.*

Assumption S2. *$\boldsymbol{D}$ and $\boldsymbol{A}_{(m)}$ $(m = 1, ..., M)$ are positive definite matrices.*

Assumption S1 is essential to obtain the influence function which is commonly used in the robust estimation literature such as Simpson et al. (1992), Coakley and Hettamansperger (1993) and Mathieu (2022). Assumption S2 is mild and can be easily satisfied.

Theorem S1. *Suppose that Assumptions S1 and S2 hold. Then, the influence function of the RJMA estimator in the case where the candidate models are not necessarily nested has the following form:*

$$\mathrm{IF}\left(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)}\right) = \sum_{m=1}^{M-1} \frac{\partial}{\partial \epsilon} w_{\epsilon,m}^* \Big|_{\epsilon=0} (\widetilde{\Theta}_{(m)} - \widetilde{\Theta}_{(M)}) + \sum_{m=1}^{M} w_m^* \mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)}), \tag{S1.2}$$

where $\mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)})$ *is a* p*-dimensional vector which contains* $p - k_m$ *zeroes and all the elements of* $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})$ *with* $\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)}) = \boldsymbol{A}_{(m)}^{-1} \boldsymbol{R}_{(m)}$ *and* $\boldsymbol{R}_{(m)} = h(\boldsymbol{x}_{(m)}^*) \psi(y^* - \boldsymbol{x}_{(m)}^{*T} \breve{\Theta}_{(m)}) \boldsymbol{x}_{(m)}^*$ *for* $m = 1, ..., M$*, and* $(\frac{\partial}{\partial \epsilon} w_{\epsilon,1}^*, ..., \frac{\partial}{\partial \epsilon} w_{\epsilon,M-1}^*)^T|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} \boldsymbol{\pi}_\epsilon^*|_{\epsilon=0} = \boldsymbol{D}^{-1}(-\boldsymbol{J} + \boldsymbol{K} + \boldsymbol{L})$ *with*

$$\boldsymbol{J} = \int h(\boldsymbol{x}) \rho_2 \left(y - \boldsymbol{x}^T \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)} \right) \boldsymbol{B} \boldsymbol{x}^T \sum_{m=1}^{M} w_m^* \mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)}) d\mathrm{G},$$

$$\boldsymbol{K} = \int h(\boldsymbol{x}) \psi \left(y - \boldsymbol{x}^T \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)} \right) \boldsymbol{H} d\mathrm{G}$$

and

$$\boldsymbol{L} = h(\boldsymbol{x}^*)\rho_2\left(y^* - \boldsymbol{x}^{*T}\sum_{m=1}^{M} w_m^*\widetilde{\Theta}_{(m)}\right)\boldsymbol{B}^*.$$

Proof. The proof is the same as that of Theorem S1 in Wang et al. (2024). So we omit it here. □

Theorem S1 derives the expression of influence function for the model averaging estimator which includes two parts, one is $\frac{\partial}{\partial\epsilon}\boldsymbol{\pi}_\epsilon^*|_{\epsilon=0}$ which measures the stability of the weight estimator given an infinitesimal contamination, and the other is the influence function of each candidate parameter estimator, i.e., $\mathrm{IF}(\boldsymbol{z};\breve{\Theta}_{(m)})$, $m = 1, ..., M$. Based on (S1.2), one can define the influence of an outlier $\boldsymbol{z}$ on the predicted value at $\boldsymbol{x}_0$ as

$$\boldsymbol{x}_0^T\mathrm{IF}\left(\boldsymbol{z};\sum_{m=1}^{M} w_m^*\widetilde{\Theta}_{(m)}\right), \tag{S1.3}$$

where $\boldsymbol{x}_0$ comes from the distribution G (see also Hampel et al. (1986)).

In the following, we shall show that under some regularity conditions, the proposed model averaging estimator is robust in terms that it has a bounded influence function.

Assumption S3. *$\rho_2(x)$ is bounded for $x \in \mathbb{R}$ almost surely.*

Assumption S4. *$|\boldsymbol{x}^T(\widetilde{\Theta}_{(m)} - \widetilde{\Theta}_{(M)})|$ and $\|\breve{\Theta}_{(m)}\|$ are bounded almost surely for $m = 1, ..., M$.*

Assumption S3 is often used to ensure that the robust estimator has a bounded influence function (Simpson et al., 1992; Coakley and Hettamansperger, 1993). Specially, when the loss function ρ is chosen to be Huber's function, it is clear that $\rho_2(x)$ is bounded. Since $\boldsymbol{x}^T\widetilde{\Theta}_{(m)}$, $m = 1, ..., M$, are all the estimators of the conditional mean $\mathrm{E}(y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\Theta}$ with $\boldsymbol{\Theta} = (\theta_1, ..., \theta_p)^T$, the requirement that $|\boldsymbol{x}^T(\widetilde{\Theta}_{(m)} - \widetilde{\Theta}_{(M)})|$ is bounded almost surely is mild and can be easily satisfied. For the regression model, it is reasonable to require that $\boldsymbol{\Theta} \in \mathcal{G}$, where $\mathcal{G}$ is a compact subset of $\mathbb{R}^p$ (Feng et al., 2021; Pan et al., 2025). So, in the case where p is fixed, $\|\boldsymbol{\Theta}\|$ is bounded. When p tends to infinity as the sample size increases, if $\boldsymbol{\Theta}$ is sparse, i.e., the most elements of $\boldsymbol{\Theta}$ are close to zero, then $\|\boldsymbol{\Theta}\|$ is also bounded. For example, in the simulation studies of Hansen (2007) and Hansen and Racine (2012), θ_j has the form of $c(2\alpha)^{1/2}j^{-\alpha-1/2}$ with c being a positive constant and $\alpha = 0.5, 1$ and 1.5, respectively. It is clear that $\|\boldsymbol{\Theta}\|^2 = 2c^2\alpha\sum_{j=1}^{p} 1/j^{2\alpha+1}$ is bounded even if p tends to infinity as $n \to \infty$. On the other hand, if we conduct a model screening procedure to select M good candidate models, then the requirement that $\boldsymbol{x}_{(m)}^T\Theta_{(m)}$ $(m = 1, ..., M)$ are close to $\boldsymbol{x}^T\boldsymbol{\Theta}$ is plausible, where $\Theta_{(m)}$ is the parameter vector corresponding to the m^{th} candidate model. Hence, by the definition of $\breve{\Theta}_{(m)}$, $\|\breve{\Theta}_{(m)}\|$ is often close to $\|\boldsymbol{\Theta}\|$. Consequently, the second part of

Assumption S4 is reasonable as well.

Theorem S2. *Suppose that Assumption 1, the second part of Assumption 5 and Assumptions S1-S4 hold. If M is fixed, then* $\|\mathrm{IF}(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)})\|$ *is bounded.*

Proof. See Section S9 of the Supplementary Material. □

Theorem S2 demonstrates the robustness of the proposed optimal model averaging estimator by showing that its influence function is bounded.

S1.2 Empirical prediction influence function

Since the cumulative distribution function G is often unknown, we cannot obtain $\widetilde{\Theta}_{(m)}$ and $\boldsymbol{w}^*$ directly. One may replace G with its empirical version in (S1.2). However, the complexity of the closed-form expression of $\mathrm{IF}(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)})$ makes the influence function not applicable in the practical situation. Hence, similar to Wang et al. (2024), we define the empirical prediction influence function (EPIF) to measure the effect of an outlier on the robust model averaging and selection methods quantitatively.

Let G_{n-1} denote the empirical distribution function associated with a dataset $\mathcal{E} = (y_i, \boldsymbol{x}_i)_{i=1}^{n-1}$ from G. Denote $\mathcal{E}^*$ as a new dataset which has one more fixed point $\boldsymbol{z} = (y^*, \boldsymbol{x}^*)$ than $\mathcal{E}$. Based on $\mathcal{E}^*$, we construct an empirical distribution function G_n^*. Under the m^{th} candidate model, we

define $\overline{\Theta}^*_{(m)}$ and $\overline{\Theta}_{(m)}$ as the solutions which minimize $Q_n(\Theta_{(m)})$ based on the datasets $\mathcal{E}^*$ and $\mathcal{E}$, respectively. Denote $\check{\Theta}^*_{(m)}$ and $\check{\Theta}_{(m)}$ as two p-dimensional vectors which consist of all the elements of $\overline{\Theta}^*_{(m)}$ and $\overline{\Theta}_{(m)}$, respectively, and $p - k_m$ zeroes. We then utilize the proposed RJMA criterion to choose weight estimators based on the datasets $\mathcal{E}^*$ and $\mathcal{E}$, and denote them as $\widetilde{\boldsymbol{w}}^* = (\widetilde{w}^*_1, ..., \widetilde{w}^*_M)^T$ and $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, ..., \widetilde{w}_M)^T$, respectively. It is clear that $\mathrm{G}^*_n = (1-\frac{1}{n})\mathrm{G}_{n-1}+\frac{1}{n}\Delta_{\boldsymbol{z}}$. Hence, similar to Wang et al. (2024), we define the empirical influence function of the model averaging estimator $\sum_{m=1}^M \widetilde{w}_m \check{\Theta}_{(m)}$ at the point $\boldsymbol{z} = (y^*, \boldsymbol{x}^*)$ as

$$\mathrm{EIF}(\boldsymbol{z}) = n\left[\sum_{m=1}^M \widetilde{w}^*_m \check{\Theta}^*_{(m)} - \sum_{m=1}^M \widetilde{w}_m \check{\Theta}_{(m)}\right]. \tag{S1.4}$$

Note that the EIF is proportional to the difference between the estimates based on the distributions G and G_ϵ, and can be regarded as the empirical version of (S1.1). So the EPIF (the empirical version of (S1.3)) has the form of

$$\mathrm{EPIF}(\boldsymbol{z}) = n\boldsymbol{x}_0^T\left[\sum_{m=1}^M \widetilde{w}^*_m \check{\Theta}^*_{(m)} - \sum_{m=1}^M \widetilde{w}_m \check{\Theta}_{(m)}\right], \tag{S1.5}$$

where $\boldsymbol{x}_0$ comes from the distribution G and is independent of the dataset $\mathcal{E}$.

The robust jackknife model selection estimators corresponding to our proposed RJMA estimators RJMA_A and RJMA_H are denoted as CV_A and

CV_H, respectively. Without loss of generality, we let the k^{th} and s^{th} candidate models be selected by the robust jackknife model selection method based on the datasets $\mathcal{E}^*$ and $\mathcal{E}$, respectively. Hence, by letting $\widetilde{w}_k^* = 1$ and $\widetilde{w}_s = 1$, (S1.5) degenerates into

$$\mathrm{EPIF}^*(\boldsymbol{z}) = n\boldsymbol{x}_0^T \left[\check{\Theta}_{(k)}^* - \check{\Theta}_{(s)}\right], \tag{S1.6}$$

which can be regarded as the EPIF of the robust jackknife model selection method. In fact, (S1.6) is the difference between the two robust jackknife model selection predictions based on the different datasets $\mathcal{E}^*$ and $\mathcal{E}$.

In the following, we examine the EPIFs of the proposed robust model averaging and the corresponding model selection estimators through simulation. To measure the quantitative robustness of RJMA_A, RJMA_H, CV_A and CV_H, we calculate the average absolute EPIFs of different estimators at the point $\boldsymbol{z}$ based on K replications,

$$I(\boldsymbol{z}) = \frac{1}{K}\sum_{k=1}^{K}\left|\mathrm{EPIF}^{(k)}(\boldsymbol{z})\right|,$$

where $\mathrm{EPIF}^{(k)}(\boldsymbol{z})$ is the empirical prediction influence function at the point $\boldsymbol{z}$ from the k^{th} replication based on different criteria. Note that if a model averaging or model selection estimator is more robust to outliers, then it will have a smaller value of $I(\boldsymbol{z})$.

Example 1: The true data generating process is the same as that in Case

1 of Simulation 1 in Section 4. In this case, all the candidate models are misspecified.

Example 2: The true data generating process follows

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{6} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, ..., n,$$

where $x_{i1} = 1$ and the remaining x_{ij} are independent and identically distributed as $\mathcal{N}(0, 1)$, $\boldsymbol{\beta} = (0.5, 0.2, 1.2, 0.01, 0, 0.3)^T$, and the error term ε_i is also $\mathcal{N}(0, 1)$ and independent of the covariates $\boldsymbol{x}_i$. Here, we consider a non-nested setup and so there are totally $M = 2^5 - 1$ candidate models with each containing the intercept term. In this case, the true model is included in the candidate model set.

For Examples 1 and 2, the sample sizes of $\mathcal{E}$ are taken as $n = 50$ and 100, respectively. The dataset $\mathcal{E}^*$ is obtained by adding one outlier point $\boldsymbol{z} = (100, \boldsymbol{x}^{*T})$, where $\boldsymbol{x}^* = (x_1^*, ..., x_p^*)^T$ with $x_1^* = 1$ and $x_2^* = \cdots = x_p^* = 5$. We set p to be 100 and 6 in Examples 1 and 2, respectively. The number of replications is $K = 200$. The simulation results for Examples 1 and 2 are summarized in Tables S1 and S2, respectively.

It is clear from Table S1 that when the sample size $n = 50$, RJMA$_H$ usually has the best performance and RJMA$_A$ is superior to CV$_A$. When the sample size increases to $n = 100$, RJMA$_H$ is the most favored estimator in terms of minimizing $I(z)$. For RJMA$_A$, its values of $I(z)$ are larger than

Table S1: $I(\boldsymbol{z})$ values of different methods for Example 1

n	R^2	RJMA$_A$	CV$_A$	RJMA$_H$	CV$_H$	n	R^2	RJMA$_A$	CV$_A$	RJMA$_H$	CV$_H$
	0.1	10.371	11.579	**5.526**	6.845		0.1	13.821	14.409	**4.861**	8.688
	0.3	11.085	16.293	**6.374**	9.455		0.3	15.380	16.791	**7.182**	12.182
50	0.5	13.476	16.381	**7.865**	11.742	100	0.5	19.252	22.632	**10.466**	16.441
	0.7	16.077	20.274	**8.721**	18.990		0.7	20.497	31.607	**10.570**	26.435
	0.9	18.499	29.703	**12.274**	27.851		0.9	22.943	35.122	**12.844**	34.354

Table S2: $I(\boldsymbol{z})$ values of different methods for Example 2

n	RJMA$_A$	CV$_A$	RJMA$_H$	CV$_H$	n	RJMA$_A$	CV$_A$	RJMA$_H$	CV$_H$
50	5.896	8.235	**3.440**	9.356	100	6.639	10.742	**2.594**	8.935

those of RJMA$_H$ and CV$_H$ for the cases where $R^2 = 0.1$, 0.3 and 0.5. However, when $R^2 = 0.7$ and 0.9, RJMA$_A$ performs only worse than RJMA$_H$. From Table S2, we see that when the true model is in the candidate model set, RJMA$_A$ and RJMA$_H$ perform better than CV$_A$ and CV$_H$ in terms of minimizing $I(z)$, respectively. In summary, we conclude that our proposed RJMA methods especially RJMA$_H$ can yield a more reliable result in the presence of outliers.

S2. Technical lemma

Let $Q_{ni}(\Theta_{(m)}) = \sum_{j\neq i}^{n} h(\boldsymbol{x}_{j(m)})\rho(y_j - \boldsymbol{x}_{j(m)}^T\Theta_{(m)})$. To prove the theorems, we need the following lemma.

Lemma 1. *Suppose that Assumptions 1-6 hold. Then*

(i) $\max_{1\le m\le M}\|\hat{\Theta}_{(m)}-\Theta^*_{(m)}\|=O_p(M^{1/2}n^{-1/2})$;

(ii) $\max_{1\le i\le n}\max_{1\le m\le M}\|\hat{\Theta}^{[-i]}_{(m)}-\Theta^*_{(m)}\|=O_p(M^{1/2}n^{-1/2})$.

Proof. (i) Let $\alpha_n=M^{1/2}n^{-1/2}$. Following Fan and Peng (2004), to verify Part (i) of Lemma 1, it suffices to show that for any given $\epsilon>0$, there exists a sufficiently large constant C_0 such that for large n, we have

$$\mathrm{P}\left\{\inf_{1\le m\le M}\inf_{\|\boldsymbol{u}_m\|=C_0}[Q_n(\Theta^*_{(m)}+\alpha_n\boldsymbol{u}_m)-Q_n(\Theta^*_{(m)})]>0\right\}\ge 1-\epsilon, \quad \text{(S2.7)}$$

which implies that with probability at least $1-\epsilon$, for the m^{th} model, there is a local minimizer $\hat{\Theta}_{(m)}$ in the ball $\{\Theta^*_{(m)}+\alpha_n\boldsymbol{u}_m:\|\boldsymbol{u}_m\|\le C_0\}$ such that $\max_{1\le m\le M}\|\hat{\Theta}_{(m)}-\Theta^*_{(m)}\|=O_p(\alpha_n)$. It is also a global minimizer by the convexity of $Q_n(\Theta_{(m)})$.

Denote $G_{inm}=h(\boldsymbol{x}_{i(m)})\{\rho(\varepsilon^*_{i(m)}-\alpha_n\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m)-\rho(\varepsilon^*_{i(m)})+\alpha_n\psi(\varepsilon^*_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m\}$, $H_{nm}=\sum_{i=1}^n\mathrm{E}\left[G_{inm}\right]$, and $D_{nm}(\Theta^*_{(m)})=\sum_{i=1}^n\psi(\varepsilon^*_{i(m)})h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}$. Then, some elementary calculations yield that

$$\begin{aligned}
&Q_n(\Theta^*_{(m)}+\alpha_n\boldsymbol{u}_m)-Q_n(\Theta^*_{(m)})\\
=&\sum_{i=1}^n h(\boldsymbol{x}_{i(m)})\rho\left\{y_i-\boldsymbol{x}^T_{i(m)}(\Theta^*_{(m)}+\alpha_n\boldsymbol{u}_m)\right\}-\sum_{i=1}^n h(\boldsymbol{x}_{i(m)})\rho(y_i-\boldsymbol{x}_{i(m)}\Theta^*_{(m)})\\
=&H_{nm}-\alpha_n D^T_{nm}(\Theta^*_{(m)})\boldsymbol{u}_m+\sum_{i=1}^n\{G_{inm}-\mathrm{E}\left[G_{inm}\right]\}\\
\equiv&\Lambda_{1n(m)}+\Lambda_{2n(m)}+\Lambda_{3n(m)}.
\end{aligned}$$

In the following, we prove that $\Lambda_{1n(m)}$ asymptotically dominates $\Lambda_{2n(m)}$ and $\Lambda_{3n(m)}$ uniformly in m.

We first consider $\Lambda_{1n(m)}$. It is clear that

$$G_{inm} = \int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m} h(\boldsymbol{x}_{i(m)}) \left\{ \psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*) \right\} dt. \tag{S2.8}$$

By Fubini's theorem and Taylor's expansion, we obtain

$$\begin{aligned}
&\sum_{i=1}^n \mathrm{E}\left[G_{inm}|X\right] \\
=&\sum_{i=1}^n h(\boldsymbol{x}_{i(m)}) \mathrm{E}\left[\int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m} \left\{ \psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*) \right\} dt | X \right] \\
=&\sum_{i=1}^n h(\boldsymbol{x}_{i(m)}) \int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m} \left\{ \mathrm{E}[\psi(\varepsilon_{i(m)}^* + t)|X] - \mathrm{E}[\psi(\varepsilon_{i(m)}^*)|X] \right\} dt \\
=&\sum_{i=1}^n h(\boldsymbol{x}_{i(m)}) \int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m} \left\{ R_i\left(\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* + t\right) - R_i\left(\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*\right) \right\} dt \\
=&\sum_{i=1}^n h(\boldsymbol{x}_{i(m)}) \int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m} R_{i1}\left(\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* + \xi_i(t)\right) t dt,
\end{aligned} \tag{S2.9}$$

where $\xi_i(t)$ lies between 0 and $-\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m$. From Assumption 6, it is seen that

$$\max_{1\le m\le M} \max_{1\le i\le n} |\alpha_n \boldsymbol{x}_{i(m)}^T \boldsymbol{u}_m| \le \alpha_n \max_{1\le m\le M} \max_{1\le i\le n} \|\boldsymbol{x}_{i(m)}\| \|\boldsymbol{u}_m\| = o(1) \tag{S2.10}$$

almost surely. Hence, it follows from Assumptions 3 and 4 that

$$\sum_{i=1}^n \mathrm{E}\left[G_{inm}|X\right] \ge \frac{c}{2} \alpha_n^2 \boldsymbol{u}_m^T \left(\sum_{i=1}^n h(\boldsymbol{x}_{i(m)}) \boldsymbol{x}_{i(m)} \boldsymbol{x}_{i(m)}^T \right) \boldsymbol{u}_m \ge \frac{c}{2} n \alpha_n^2 C_0^2$$

almost surely, which indicates that

$$\Lambda_{1n(m)} \ge \frac{c}{2} n \alpha_n^2 C_0^2. \tag{S2.11}$$

We now consider $\Lambda_{2n(m)}$. Denote

$$\omega_{inm} = \psi(\varepsilon^*_{i(m)})h(\boldsymbol{x}_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m - \mathrm{E}[h(\boldsymbol{x}_{i(m)})\psi(\varepsilon^*_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m].$$

Note that $\mathrm{E}(\omega_{inm}) = 0$ and ω_{inm} $(i = 1, ..., n)$ are mutually independent. Thus, invoking Assumptions 1 and 5, we obtain

$$\begin{aligned}
\mathrm{E}\left\{\max_{1\le m\le M}\left(\sum_{i=1}^{n}\omega_{inm}\right)^2\right\} &\le \sum_{m=1}^{M}\mathrm{E}\left\{\left(\sum_{i=1}^{n}\omega_{inm}\right)^2\right\} \\
&= \sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{Var}\left\{h(\boldsymbol{x}_{i(m)})\psi(\varepsilon^*_{i(m)})\boldsymbol{u}_m^T\boldsymbol{x}_{i(m)}\right\} \\
&\le C\sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{E}\left\{\left(h(\boldsymbol{x}_{i(m)})\boldsymbol{u}_m^T\boldsymbol{x}_{i(m)}\right)^2\right\} \\
&\le C\sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{E}(\|h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}\|^2\|\boldsymbol{u}_m\|^2) \\
&= O(nM)C_0^2.
\end{aligned}$$

So $\max_{1\le m\le M}\sum_{i=1}^{n}\omega_{inm} = O_p(n^{1/2}M^{1/2})C_0$. By the definition of $\Theta^*_{(m)}$, we obtain

$$\sum_{i=1}^{n}\mathrm{E}\{h(\boldsymbol{x}_{i(m)})\psi(\varepsilon^*_{i(m)})\boldsymbol{x}_{i(m)}\} = 0.$$

Hence, it follows that

$$\max_{1\le m\le M}\sum_{i=1}^{n}h(\boldsymbol{x}_{i(m)})\psi(\varepsilon^*_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m = O_p(n^{1/2}M^{1/2})C_0.$$

This indicates that

$$\max_{1\le m\le M}|\Lambda_{2n(m)}| = O_p(\alpha_n n^{1/2}M^{1/2})C_0 = O_p(n\alpha_n^2)C_0. \tag{S2.12}$$

Finally, we consider $\Lambda_{3n(m)}$. Let $\gamma_{inm} = G_{inm} - \mathrm{E}\left[G_{inm}\right]$. It is obvious that $\mathrm{E}\left[\gamma_{inm}\right] = 0$ and γ_{inm} $(i = 1, ..., n)$ are mutually independent. Hence, we obtain

$$\begin{aligned}
\mathrm{E}\left[\max_{1\leq m\leq M}\left(\sum_{i=1}^{n}\gamma_{inm}\right)^2 |X\right] &\leq \sum_{m=1}^{M}\mathrm{E}\left[\left(\sum_{i=1}^{n}\gamma_{inm}\right)^2 |X\right] \\
&= \sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{E}\left[\gamma_{inm}^2|X\right] \\
&= \sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{Var}(G_{inm}|X).
\end{aligned}$$

From Assumptions 2 and 5, Cauchy-Schwarz inequality and Fubini's theorem, we see that

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{Var}(G_{inm}|X) \\
=&\sum_{m=1}^{M}\sum_{i=1}^{n}\mathrm{Var}\left[\int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m} h(\boldsymbol{x}_{i(m)})\left\{\psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*)\right\}dt\Big|X\right] \\
\leq&\sum_{m=1}^{M}\sum_{i=1}^{n}h^2(\boldsymbol{x}_{i(m)})\mathrm{E}\left[\left\{\int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m}\left\{\psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*)\right\}dt\right\}^2\Big|X\right] \\
\leq&\sum_{m=1}^{M}\sum_{i=1}^{n}h^2(\boldsymbol{x}_{i(m)})\mathrm{E}\left[|\alpha_n\boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m|\int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m}\left\{\psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*)\right\}^2 dt\Big|X\right] \\
=&\sum_{m=1}^{M}\sum_{i=1}^{n}h^2(\boldsymbol{x}_{i(m)})|\alpha_n\boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m|\int_0^{-\alpha_n \boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m}\mathrm{E}\left[\left\{\psi(\varepsilon_{i(m)}^* + t) - \psi(\varepsilon_{i(m)}^*)\right\}^2|X\right]dt \\
=&\sum_{m=1}^{M}\sum_{i=1}^{n}(\alpha_n h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m)^2 o(1) \\
=&o(\alpha_n^2 nM)C_0^2 \qquad \text{(S2.13)}
\end{aligned}$$

almost surely, where o(1) in the third equality is unrelated to i and m because of Assumption 2 and (S2.10). Since the last equality of (S2.13) is unrelated to X, it follows that $\mathrm{E}[\max_{1\le m\le M}\Lambda^2_{3n(m)}] = o(\alpha_n^2 nM)C_0^2$. Thus, by Markov's inequality, we see that for any $\varepsilon > 0$,

$$\begin{aligned}&\mathrm{P}\left(\max_{1\le m\le M}|\Lambda_{3n(m)}| > n\alpha_n^2 C_0\varepsilon\right)\\ \le&\frac{\mathrm{E}[\max_{1\le m\le M}\Lambda^2_{3n(m)}]}{n^2\alpha_n^4 C_0^2\varepsilon^2}\\ =&o(1).\end{aligned}$$

This yields that

$$\max_{1\le m\le M}|\Lambda_{3n(m)}| = o_p(n\alpha_n^2)C_0. \tag{S2.14}$$

Combining (S2.11), (S2.12) and (S2.14), it is seen that $\Lambda_{2n(m)}$ and $\Lambda_{3n(m)}$ are both asymptotically dominated by $\Lambda_{1n(m)}$ uniformly in m when C_0 is a sufficiently large constant. Since $\Lambda_{1n(m)}$ is positive, we conclude that (S2.7) is true and this completes the proof of Part (i) of Lemma 1.

(ii) Since the proof of Part (ii) is similar to that of Part (i), we use the same notations as in Part (i) for convenience. Here, we need to show that for any $\epsilon > 0$, there exists a large constant C_0 such that for large n, we have

$$\mathrm{P}\left\{\inf_{1\le m\le M}\inf_{1\le i\le n}\inf_{\|\boldsymbol{u}_m\|=C_0}[Q_{ni}(\Theta^*_{(m)} + \alpha_n\boldsymbol{u}_m) - Q_{ni}(\Theta^*_{(m)})] > 0\right\} \ge 1-\epsilon,$$

which implies that for each candidate model, there exists a global minimizer

$\hat{\Theta}_{(m)}^{[-i]}$ such that $\max_{1\le i\le n}\max_{1\le m\le M}\|\hat{\Theta}_{(m)}^{[-i]}-\Theta_{(m)}^{*}\|=O_p(\alpha_n)$. Observe that

$$
\begin{aligned}
&Q_{ni}(\Theta_{(m)}^{*}+\alpha_n\boldsymbol{u}_m)-Q_{ni}(\Theta_{(m)}^{*})\\
=&\sum_{j\neq i}^{n}h(\boldsymbol{x}_{j(m)})\rho\left\{y_j-\boldsymbol{x}_{j(m)}^{T}(\Theta_{(m)}^{*}+\alpha_n\boldsymbol{u}_m)\right\}-\sum_{j\neq i}^{n}h(\boldsymbol{x}_{j(m)})\rho(y_j-\boldsymbol{x}_{j(m)}\Theta_{(m)}^{*})\\
=&H_{nmi}-\alpha_n D_{nmi}^{T}(\Theta_{(m)}^{*})\boldsymbol{u}_m+\sum_{j\neq i}^{n}\left\{G_{jnm}-\mathrm{E}\left[G_{jnm}|X\right]\right\}\\
\equiv&\Lambda_{1n(m)}^{[-i]}+\Lambda_{2n(m)}^{[-i]}+\Lambda_{3n(m)}^{[-i]},
\end{aligned}
\tag{S2.15}
$$

where $H_{nmi}=\sum_{j\neq i}^{n}\mathrm{E}\left[G_{jnm}\right]$ and $D_{nmi}(\Theta_{(m)}^{*})=\sum_{j\neq i}^{n}h(\boldsymbol{x}_{j(m)})\psi(\varepsilon_{j(m)}^{*})\boldsymbol{x}_{j(m)}$. In the following, we are going to prove that $\Lambda_{2n(m)}^{[-i]}$ and $\Lambda_{3n(m)}^{[-i]}$ are asymptotically dominated by $\Lambda_{1n(m)}^{[-i]}$ uniformly in i and m.

We decompose $\Lambda_{1n(m)}^{[-i]}$ as $\Lambda_{1n(m)}^{[-i]}=\Lambda_{1n(m)}-\mathrm{E}\left[G_{inm}\right]$. Then we obtain

$$
\min_{1\le m\le M}\min_{1\le i\le n}\Lambda_{1n(m)}^{[-i]}\ge\min_{1\le m\le M}\Lambda_{1n(m)}-\max_{1\le m\le M}\max_{1\le i\le n}\left|\mathrm{E}\left[G_{inm}\right]\right|. \tag{S2.16}
$$

Invoking Assumptions 1 and 5, we have

$$
\begin{aligned}
&\max_{1\le m\le M}\max_{1\le i\le n}\left|\mathrm{E}\left[G_{inm}\right]\right|\\
\le&\max_{1\le m\le M}\max_{1\le i\le n}\mathrm{E}\left[h(\boldsymbol{x}_{i(m)})\int_{0}^{-\alpha_n\boldsymbol{x}_{i(m)}^{T}\boldsymbol{u}_m}|\psi(\varepsilon_{i(m)}^{*}+t)-\psi(\varepsilon_{i(m)}^{*})|dt\right]\\
\le&C\alpha_n\max_{1\le m\le M}\max_{1\le i\le n}\mathrm{E}(|h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}^{T}\boldsymbol{u}_m|)\\
\le&C\alpha_n\max_{1\le m\le M}\max_{1\le i\le n}\mathrm{E}(\|h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}\|\|\boldsymbol{u}_m\|)\\
=&O(\alpha_n)=o(1).
\end{aligned}
\tag{S2.17}
$$

Hence, it is seen from (S2.11) and (S2.17) that

$$\min_{1\le m\le M}\min_{1\le i\le n}\Lambda^{[-i]}_{1n(m)} \ge \frac{c}{2}n\alpha_n^2 C_0^2 + o(1). \tag{S2.18}$$

As for $D^T_{nmi}(\Theta^*_{(m)})\boldsymbol{u}_m$, we have the following decomposition

$$D^T_{nmi}(\Theta^*_{(m)})\boldsymbol{u}_m = D^T_{nm}(\Theta^*_{(m)})\boldsymbol{u}_m - \psi(\varepsilon^*_{i(m)})h(\boldsymbol{x}_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m.$$

So it is clear that

$$\begin{aligned}\max_{1\le m\le M}\max_{1\le i\le n}|\Lambda^{[-i]}_{2n(m)}| \le& \alpha_n \max_{1\le m\le M}|D^T_{nm}(\Theta^*_{(m)})\boldsymbol{u}_m| \\ &+ \max_{1\le m\le M}\max_{1\le i\le n}\alpha_n|\psi(\varepsilon^*_{i(m)})h(\boldsymbol{x}_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m|.\end{aligned}$$

By Assumptions 1 and 5, it is seen that

$$\max_{1\le m\le M}\max_{1\le i\le n}\alpha_n|\psi(\varepsilon^*_{i(m)})h(\boldsymbol{x}_{i(m)})\boldsymbol{x}^T_{i(m)}\boldsymbol{u}_m| = O(\alpha_n) = o(1)$$

almost surely. Hence, according to (S2.12), we have

$$\max_{1\le m\le M}\max_{1\le i\le n}|\Lambda^{[-i]}_{2n(m)}| = O_p(n\alpha_n^2)C_0. \tag{S2.19}$$

We now consider $\Lambda^{[-i]}_{3n(m)}$, which can be decomposed as

$$\Lambda^{[-i]}_{3n(m)} = \Lambda_{3n(m)} - \{G_{inm} - \mathrm{E}\,[G_{inm}]\}.$$

Applying Triangle inequality, we obtain

$$\begin{aligned}\max_{1\le m\le M}\max_{1\le i\le n}|\Lambda^{[-i]}_{3n(m)}| \le& \max_{1\le m\le M}|\Lambda_{3n(m)}| + \max_{1\le m\le M}\max_{1\le i\le n}|G_{inm}| \\ &+ \max_{1\le m\le M}\max_{1\le i\le n}|\mathrm{E}\,[G_{inm}]\,|.\end{aligned} \tag{S2.20}$$

According to (S2.8) and Assumptions 1 and 5, we have

$$\begin{aligned}
&\max_{1\le m\le M}\max_{1\le i\le n}|G_{inm}| \\
\le&\max_{1\le m\le M}\max_{1\le i\le n}h(\boldsymbol{x}_{i(m)})\int_0^{-\alpha_n\boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m}\left|\psi(\varepsilon^*_{i(m)}+t)-\psi(\varepsilon^*_{i(m)})\right|dt \\
\le&C\alpha_n\max_{1\le m\le M}\max_{1\le i\le n}|h(\boldsymbol{x}_{i(m)})\boldsymbol{x}_{i(m)}^T\boldsymbol{u}_m| \\
=&o(1) \qquad\qquad (S2.21)
\end{aligned}$$

almost surely. Hence, it follows from (S2.14), (S2.17), (S2.20) and (S2.21) that

$$\max_{1\le m\le M}\max_{1\le i\le n}|\Lambda^{[-i]}_{3n(m)}|=o_p(n\alpha_n^2)C_0. \qquad (S2.22)$$

Combining (S2.18), (S2.19) and (S2.22), it is seen that when C_0 is chosen to be a sufficiently large constant, $\Lambda^{[-i]}_{2n(m)}$ and $\Lambda^{[-i]}_{3n(m)}$ are asymptotically dominated by $\Lambda^{[-i]}_{1n(m)}$ uniformly in i and m. As $\Lambda^{[-i]}_{1n(m)}$ is positive, this completes the proof of Part (ii) of Lemma 1. □

S3. Proof of Theorem 1

Let

$$\mathrm{RCV}_n^*(\boldsymbol{w})=\mathrm{RCV}_n(\boldsymbol{w})-\sum_{i=1}^n\left[h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)-\mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)|X\right\}\right]. \qquad (S3.23)$$

Since the second term on the right-hand side of (S3.23) is independent of $\boldsymbol{w}$, minimizing $\mathrm{RCV}_n(\boldsymbol{w})$ is equivalent to minimizing $\mathrm{RCV}_n^*(\boldsymbol{w})$. As shown

by Theorem 2.1 of Li (1987), to verify Theorem 1, it suffices to show that

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{RCV}_n^*(\boldsymbol{w})-\mathrm{FPE}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|=o_p(1). \tag{S3.24}$$

By Triangle inequality, we obtain

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{RCV}_n^*(\boldsymbol{w})-\mathrm{FPE}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|\leq\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{RCV}_n^*(\boldsymbol{w})-\widetilde{\mathrm{FPE}}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|+\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{FPE}_n(\boldsymbol{w})-\widetilde{\mathrm{FPE}}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|,$$

where $\widetilde{\mathrm{FPE}}_n(\boldsymbol{w})=\sum_{i=1}^n \mathrm{E}\left[h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\}|X\right]$. Thus, we show (S3.24) by proving that

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{RCV}_n^*(\boldsymbol{w})-\widetilde{\mathrm{FPE}}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|=o_p(1) \tag{S3.25}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\mathrm{FPE}_n(\boldsymbol{w})-\widetilde{\mathrm{FPE}}_n(\boldsymbol{w})}{\mathrm{FPE}_n(\boldsymbol{w})}\right|=o_p(1). \tag{S3.26}$$

Let us first consider (S3.25). In the following, we will use the following equation repeatedly:

$$\rho(u+v)-\rho(u)=v\psi(u)+\int_0^v\{\psi(u+t)-\psi(u)\}dt. \tag{S3.27}$$

Note that (S3.27) can be regarded as the extension of Knight's identity. By

(S3.27), we obtain

$$
\begin{aligned}
&\mathrm{FPE}_n(\boldsymbol{w}) - \widetilde{\mathrm{FPE}}_n(\boldsymbol{w}) \\
=&\sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i)\rho\left(\widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) + \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \hat{\Theta}_{(m)}) \right) - h(\breve{\boldsymbol{x}}_i)\rho\left\{ \widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) \right\} \Big| \mathcal{D}_n \right\} \\
=&\sum_{i=1}^{n} \mathrm{E}\left\{ \sum_{m=1}^{M} w_m h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \hat{\Theta}_{(m)}) \psi\{ \widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) \} \Big| \mathcal{D}_n \right\} \\
&+ \sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i) \int_0^{\sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \hat{\Theta}_{(m)})} \left[\psi\{ \widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) + t \} - \psi\{ \widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) \} \right] dt \Big| \mathcal{D}_n \right\} \\
\equiv&\Pi_{n,1}(\boldsymbol{w}) + \Pi_{n,2}(\boldsymbol{w}).
\end{aligned}
$$

From Assumptions 1 and 5, Cauchy-Schwarz inequality and Lemma 1, we have

$$
\begin{aligned}
\sup_{\boldsymbol{w}\in\mathcal{W}} |\Pi_{n,1}(\boldsymbol{w})| &\le C \max_{1\le m\le M} \|\Theta_{(m)}^* - \hat{\Theta}_{(m)}\| \sum_{i=1}^{n} \max_{1\le m\le M} \|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\| \\
&= O_p(M^{1/2} n^{1/2}). \qquad \text{(S3.28)}
\end{aligned}
$$

In view of the fact that $|\psi\{\widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w}) + t\} - \psi\{\widetilde{\varepsilon}_i + \xi_i^*(\boldsymbol{w})\}| \le C$, we can prove that

$$
\sup_{\boldsymbol{w}\in\mathcal{W}} |\Pi_{n,2}(\boldsymbol{w})| = O_p(M^{1/2} n^{1/2}). \qquad \text{(S3.29)}
$$

Hence, we conclude from (S3.28) and (S3.29) that

$$
\sup_{\boldsymbol{w}\in\mathcal{W}} |\mathrm{FPE}_n(\boldsymbol{w}) - \widetilde{\mathrm{FPE}}_n(\boldsymbol{w})| = O_p(M^{1/2} n^{1/2}). \qquad \text{(S3.30)}
$$

Let $D_i(t) = \mathrm{E}\{\rho(\varepsilon_i + t) - \rho(\varepsilon_i)|\boldsymbol{x}_i\}$. Since ρ is a convex function, $D_i(t)$ is also convex with respect to t. By Assumption 1, we have $\frac{\partial D_i(t)}{\partial t}|_{t=0} = \mathrm{E}\{\psi(\varepsilon_i)|\boldsymbol{x}_i\} = 0$. Thus, it is seen that $D_i(t)$ has a global minimum at $t = 0$ for $i = 1, ..., n$, which implies that

$$\begin{aligned}\inf_{\boldsymbol{w}\in\mathcal{W}} \widetilde{\mathrm{FPE}}_n(\boldsymbol{w}) &= \inf_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^{n} \mathrm{E}\left[h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\}|X\right] \\ &= \inf_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\mathrm{E}\left[\rho\left\{\varepsilon_i + \xi_i^*(\boldsymbol{w})\right\}|\boldsymbol{x}_i\right] \\ &\geq \sum_{i=1}^{n} \mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)|\boldsymbol{x}_i\right\}. \end{aligned} \tag{S3.31}$$

Hence, it follows from (S3.30) that

$$\begin{aligned}\inf_{\boldsymbol{w}\in\mathcal{W}} \mathrm{FPE}_n(\boldsymbol{w}) &\geq \inf_{\boldsymbol{w}\in\mathcal{W}} \widetilde{\mathrm{FPE}}_n(\boldsymbol{w}) - O_p(M^{1/2}n^{1/2}) \\ &\geq \sum_{i=1}^{n} \mathrm{E}\{h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)|\boldsymbol{x}_i\} - O_p(M^{1/2}n^{1/2}). \end{aligned} \tag{S3.32}$$

It is clear that $n^{-1}\sum_{i=1}^{n} \mathrm{E}\{h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)|\boldsymbol{x}_i\}$ is bounded away from zero which indicates that with probability tending to one, $n^{-1}\mathrm{FPE}_n(\boldsymbol{w})$ is bounded away from zero uniformly in $\boldsymbol{w}$. In the following, we shall prove that

$$\sup_{\boldsymbol{w}\in\mathcal{W}} n^{-1}|\mathrm{RCV}_n^*(\boldsymbol{w}) - \widetilde{\mathrm{FPE}}_n(\boldsymbol{w})| = o_p(1). \tag{S3.33}$$

By Assumption 1, we have

$$\mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\xi_i^*(\boldsymbol{w})\psi(\varepsilon_i)|X\right\} = h(\breve{\boldsymbol{x}}_i)\xi_i^*(\boldsymbol{w})\mathrm{E}\left\{\psi(\varepsilon_i)|\boldsymbol{x}_i\right\} = 0.$$

Thus, by (S3.27), we obtain the following decomposition

$$\begin{aligned}
&n^{-1}\mathrm{RCV}_n^*(\boldsymbol{w}) - n^{-1}\widetilde{\mathrm{FPE}}_n(\boldsymbol{w}) \\
=&\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}^{[-i]}\right) - \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\} - \rho(\varepsilon_i)\right] \\
&- \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\left\{h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\} - h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)\right\}|X\right] \\
\equiv& CV_{1n}(\boldsymbol{w}) + CV_{2n}(\boldsymbol{w}) + CV_{3n}(\boldsymbol{w}),
\end{aligned}$$

where

$$CV_{1n}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}^{[-i]}\right) - \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\},$$

$$CV_{2n}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\xi_i^*(\boldsymbol{w})\psi(\varepsilon_i)$$

and

$$\begin{aligned}
CV_{3n}(\boldsymbol{w}) =& \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\int_0^{\xi_i^*(\boldsymbol{w})}\left\{\psi(\varepsilon_i + t) - \psi(\varepsilon_i)\right\}dt \\
&- \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[h(\breve{\boldsymbol{x}}_i)\int_0^{\xi_i^*(\boldsymbol{w})}\left\{\psi(\varepsilon_i + t) - \psi(\varepsilon_i)\right\}dt|X\right].
\end{aligned}$$

Hence, we prove (S3.33) by showing that (i) $\sup_{\boldsymbol{w}\in\mathcal{W}}|CV_{1n}(\boldsymbol{w})| = o_p(1)$; (ii) $\sup_{\boldsymbol{w}\in\mathcal{W}}|CV_{2n}(\boldsymbol{w})| = o_p(1)$; and (iii) $\sup_{\boldsymbol{w}\in\mathcal{W}}|CV_{3n}(\boldsymbol{w})| = o_p(1)$.

We first consider (i). From (S3.27), we see that

$$\begin{aligned}
CV_{1n}(\boldsymbol{w}) =& \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\left(\varepsilon_i^*(\boldsymbol{w}) + \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]})\right) - \rho\left\{\varepsilon_i^*(\boldsymbol{w})\right\}\right] \\
=& \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]})\psi\{\varepsilon_i^*(\boldsymbol{w})\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\int_0^{\sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]})} \left[\psi\{\varepsilon_i^*(\boldsymbol{w}) + t\} - \psi\{\varepsilon_i^*(\boldsymbol{w})\}\right] dt \\
\equiv& CV_{1n,1}(\boldsymbol{w}) + CV_{1n,2}(\boldsymbol{w}).
\end{aligned}$$

By Assumptions 1 and 5, and Lemma 1, it is seen that

$$\begin{aligned}
\sup_{\boldsymbol{w}\in\mathcal{W}} |CV_{1n,1}(\boldsymbol{w})| &\le \max_{1\le i\le n}\max_{1\le m\le M} \|\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]}\| \|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\| \\
&= O_p(M^{1/2}n^{-1/2}).
\end{aligned}$$

For $CV_{1n,2}(\boldsymbol{w})$, we can similarly prove that $\sup_{\boldsymbol{w}\in\mathcal{W}} |CV_{1n,2}(\boldsymbol{w})| = O_p(M^{1/2}n^{-1/2})$. Thus, we have

$$\sup_{\boldsymbol{w}\in\mathcal{W}} |CV_{1n}(\boldsymbol{w})| = O_p(M^{1/2}n^{-1/2}) = o_p(1).$$

We then consider (ii). Denote $h_n = 1/(\overline{k}^{1/2}\log n)$. For any $\boldsymbol{w} = (w_1, ..., w_M)^T \in \mathcal{W}$ and $\overline{\boldsymbol{w}}_j = (\overline{w}_{j1}, ..., \overline{w}_{jM})^T \in \mathcal{W}$, we construct grids based on regions of the form $W_j = \{\boldsymbol{w} : |\boldsymbol{w} - \overline{\boldsymbol{w}}_j|_1 \le h_n\}$, where $|\boldsymbol{w} - \overline{\boldsymbol{w}}_j|_1 = \sum_{m=1}^{M} |w_m - \overline{w}_{jm}|$. If we select $\overline{\boldsymbol{w}}_j$ to lay on a grid, then $\mathcal{W}$ can be covered with $N = O(1/h_n^{M-1})$ regions W_j, $j = 1, ..., N$. By Assumptions 1 and 5, and Cauchy-Schwarz inequality, it is readily seen

that

$$
\begin{aligned}
&\max_{1\le j\le N}\sup_{\boldsymbol{w}\in W_j}|CV_{2n}(\boldsymbol{w})-CV_{2n}(\overline{\boldsymbol{w}}_j)|\\
=&\max_{1\le j\le N}\sup_{\boldsymbol{w}\in W_j}\left|\sum_{m=1}^{M}(\overline{w}_{jm}-w_m)\frac{1}{n}\sum_{i=1}^{n}h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\psi(\varepsilon_i)\right|\\
\le& C\max_{1\le m\le M}\|\Theta_{(m)}^*\|\max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^{n}\|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\|\max_{1\le j\le N}\sup_{\boldsymbol{w}\in W_j}\sum_{m=1}^{M}|\overline{w}_{jm}-w_m|\\
\le& CO(\overline{k}^{1/2})h_n=o(1)
\end{aligned}
$$

almost surely. Hence, it follows that

$$
\begin{aligned}
\sup_{\boldsymbol{w}\in\mathcal{W}}|CV_{2n}(\boldsymbol{w})|&\le\max_{1\le j\le N}\sup_{\boldsymbol{w}\in W_j}|CV_{2n}(\boldsymbol{w})|\\
&\le\max_{1\le j\le N}|CV_{2n}(\overline{\boldsymbol{w}}_j)|+\max_{1\le j\le N}\sup_{\boldsymbol{w}\in W_j}|CV_{2n}(\boldsymbol{w})-CV_{2n}(\overline{\boldsymbol{w}}_j)|\\
&=\max_{1\le j\le N}|CV_{2n}(\overline{\boldsymbol{w}}_j)|+o(1). \qquad \text{(S3.34)}
\end{aligned}
$$

Consequently, to prove $\sup_{\boldsymbol{w}\in\mathcal{W}}|CV_{2n}(\boldsymbol{w})|=o_p(1)$, it suffices to verify that $\max_{1\le j\le N}|CV_{2n}(\overline{\boldsymbol{w}}_j)|=o_p(1)$. Denote $u_i(\overline{\boldsymbol{w}}_j)=h(\breve{\boldsymbol{x}}_i)(\mu_i-\sum_{m=1}^{M}\overline{w}_{jm}\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*)$ and $e_n=(n^{1+\delta}M\overline{k})^{1/2}$ for some $0<\delta<1$. From Assumption 1, we have

$$
\begin{aligned}
|u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)|&=\left|\sum_{m=1}^{M}\overline{w}_{jm}h(\breve{\boldsymbol{x}}_i)(\mu_i-\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*)\psi(\varepsilon_i)\right|\\
&\le\max_{1\le m\le M}Ch(\breve{\boldsymbol{x}}_i)|\mu_i-\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*|\equiv b_i.
\end{aligned}
$$

By Boole's inequality, we obtain that for any $\epsilon > 0$,

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\le j\le N}|CV_{2n}(\overline{\boldsymbol{w}}_j)|\ge 2\epsilon\right)\\
=&\mathrm{P}\left(\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^{n}u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\right|\ge 2\epsilon\right)\\
\le&\mathrm{P}\left(\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^{n}u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\le e_n)\right|\ge \epsilon\right)\\
&+\mathrm{P}\left(\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^{n}u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\ge e_n)\right|\ge \epsilon\right)\\
\le&N\max_{1\le j\le N}\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\le e_n)\right|\ge \epsilon\right)\\
&+\mathrm{P}\left(\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^{n}u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\ge e_n)\right|\ge \epsilon\right)\\
\equiv&CV_{2n,1}+CV_{2n,2}.
\end{aligned}$$

By Cauchy-Schwarz and Jensen's inequalities, and Assumptions 1 and 5, it is seen that

$$\mathrm{E}\left\{u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\le e_n)|X\right\}=u_i(\overline{\boldsymbol{w}}_j)\mathrm{I}(b_i\le e_n)\mathrm{E}\left\{\psi(\varepsilon_i)|\boldsymbol{x}_i\right\}=0$$

and

$$\begin{aligned}
\mathrm{Var}\left\{u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i\le e_n)|X\right\}\le&C\left\{\sum_{m=1}^{M}\overline{w}_{jm}h(\breve{\boldsymbol{x}}_i)(\mu_i-\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*)\right\}^2\\
\le&C\max_{1\le m\le M}h^2(\breve{\boldsymbol{x}}_i)\left(\mu_i-\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right)^2\\
\le&C\left\{(h(\breve{\boldsymbol{x}}_i)\mu_i)^2+\max_{1\le m\le M}\|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\|^2\|\Theta_{(m)}^*\|^2\right\}\\
\le&C\overline{k},
\end{aligned}$$

which holds uniformly in i and j almost surely. Denote $\mathrm{P}_X(\cdot)$ as the conditional probability operator $\mathrm{P}(\cdot|X)$. Note that $u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \leq e_n)$ $(i = 1, ..., n)$ are mutually independent conditional on X, then by Bernstein's inequality, we obtain

$$\mathrm{P}_X\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \leq e_n)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2C\overline{k} + 2e_n\epsilon/3}\right) \tag{S3.35}$$

almost surely. Since the right-hand side of (S3.35) is unrelated to X, it is seen that

$$\begin{aligned}
&\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \leq e_n)\right| \geq \epsilon\right) \\
=&\mathrm{E}\left\{\mathrm{P}_X\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \leq e_n)\right| \geq \epsilon\right)\right\} \\
\leq&2\exp\left(-\frac{n\epsilon^2}{2C\overline{k} + 2e_n\epsilon/3}\right).
\end{aligned}$$

This yields that

$$\begin{aligned}
CV_{2n,1} &= N \max_{1\leq j\leq N} \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \leq e_n)\right| \geq \epsilon\right) \\
&\leq 2N\exp\left(-\frac{n\epsilon^2}{2C\overline{k} + 2(n^{1+\delta}M\overline{k})^{1/2}\epsilon/3}\right) \\
&= o(1),
\end{aligned} \tag{S3.36}$$

where the last equality follows from the fact that $M^3\overline{k}\log^2(\overline{k}^{1/2}\log n)/n^{1-\delta} = o(1)$. By Boole's, Markov's and Cauchy-Schwarz inequalities, Assumption

5, and the fact that

$$\left\{ \max_{1\le j\le N} \left| \frac{1}{n}\sum_{i=1}^{n} u_i(\overline{\boldsymbol{w}}_j)\psi(\varepsilon_i)\mathrm{I}(b_i \ge e_n) \right| \ge \epsilon \right\} \subset \left\{ \max_{1\le i\le n} b_i \ge e_n \right\},$$

we can show that

$$\begin{aligned} CV_{2n,2} &\le \mathrm{P}\left(\max_{1\le m\le M} \max_{1\le i\le n} Ch(\breve{\boldsymbol{x}}_i)|\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*| \ge e_n \right) \\ &\le \sum_{m=1}^{M}\sum_{i=1}^{n} \mathrm{P}\left(Ch(\breve{\boldsymbol{x}}_i)|\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*| \ge e_n \right) \\ &\le \sum_{m=1}^{M}\sum_{i=1}^{n} \frac{C}{e_n^2}\mathrm{E}\left(h^2(\breve{\boldsymbol{x}}_i)|\mu_i - \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*|^2 \right) \\ &\le \sum_{m=1}^{M}\sum_{i=1}^{n} \frac{C}{e_n^2}\mathrm{E}\left(h^2(\breve{\boldsymbol{x}}_i)\mu_i^2 + \|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\|^2 \|\Theta_{(m)}^*\|^2 \right) \\ &= o(1). \end{aligned} \tag{S3.37}$$

Combining (S3.36) and (S3.37), we have $\max_{1\le j\le N} |CV_{2n}(\overline{\boldsymbol{w}}_j)| = o_p(1)$. Thus, it follows from (S3.34) that $\sup_{\boldsymbol{w}\in\mathcal{W}} |CV_{2n}(\boldsymbol{w})| = o_p(1)$.

We now turn our attention to (iii). Similar to the treatment of $CV_{2n}(\boldsymbol{w})$, we can show that

$$\sup_{\boldsymbol{w}\in\mathcal{W}} |CV_{3n}(\boldsymbol{w})| = o_p(1).$$

So (S3.25) is correct. By (S3.30) and (S3.32), it follows that (S3.26) is also true. This completes the proof of Theorem 1.

S4. Proof of Theorem 2

Let $\alpha_n = n^{-1/4}M^{1/4}$. Following the arguments in Fan and Li (2001) and Chen et al. (2018), to verify Theorem 2, it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant C_0 such that for large n, we have

$$\mathrm{P}\left\{\inf_{\|\boldsymbol{u}\|=C_0,(\boldsymbol{w}^0+\alpha_n\boldsymbol{u})\in\mathcal{W}} \mathrm{RCV}_n(\boldsymbol{w}^0+\alpha_n\boldsymbol{u}) > \mathrm{RCV}_n(\boldsymbol{w}^0)\right\} \geq 1-\epsilon, \quad \text{(S4.38)}$$

which implies that with probability at least $1-\epsilon$, there exists a local minimizer $\hat{\boldsymbol{w}}$ in the bounded closed domain $\{\boldsymbol{w}^0+\alpha_n\boldsymbol{u} : \|\boldsymbol{u}\| \leq C_0, (\boldsymbol{w}^0+\alpha_n\boldsymbol{u}) \in \mathcal{W}\}$ such that $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^0\| = O_p(\alpha_n)$.

We shall show that $\mathrm{RCV}_n(\boldsymbol{w})$ is a convex function on the $\mathcal{W}$ which indicates that $\hat{\boldsymbol{w}}$ is also a global minimizer of $\mathrm{RCV}_n(\boldsymbol{w})$. Let $\boldsymbol{w}_1 = (w_{11}, ..., w_{1M})^T \in \mathcal{W}$, $\boldsymbol{w}_2 = (w_{21}, ..., w_{2M})^T \in \mathcal{W}$ and $\lambda \in [0,1]$. It is obvious that $\mathcal{W}$ is a convex set. To show $\mathrm{RCV}_n(\boldsymbol{w})$ is a convex function, we only need to prove that

$$\mathrm{RCV}_n(\lambda\boldsymbol{w}_1 + (1-\lambda)\boldsymbol{w}_2) \leq \lambda\mathrm{RCV}_n(\boldsymbol{w}_1) + (1-\lambda)\mathrm{RCV}_n(\boldsymbol{w}_2).$$

Denote $\widehat{h}_{i1} = y_i - \sum_{m=1}^{M} w_{1m}\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}$, $\widehat{h}_{i2} = y_i - \sum_{m=1}^{M} w_{2m}\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}$ and

$\widehat{a}_{im} = \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}^{[-i]}$. By the convexity of ρ, it is seen that

$$\begin{aligned}
&\mathrm{RCV}_n(\lambda \boldsymbol{w}_1 + (1-\lambda)\boldsymbol{w}_2) \\
=&\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \lambda\sum_{m=1}^{M} w_{1m}\widehat{a}_{im} - (1-\lambda)\sum_{m=1}^{M} w_{2m}\widehat{a}_{im}\right) \\
=&\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(\lambda h_{i1} + (1-\lambda)h_{i2}\right) \\
\leq&\lambda\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho(\widehat{h}_{i1}) + (1-\lambda)\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho(\widehat{h}_{i2}) \\
=&\lambda\mathrm{RCV}_n(\boldsymbol{w}_1) + (1-\lambda)\mathrm{RCV}_n(\boldsymbol{w}_2).
\end{aligned}$$

Hence, $\mathrm{RCV}_n(\boldsymbol{w})$ is a convex function.

Observe that

$$\begin{aligned}
&\mathrm{RCV}_n(\boldsymbol{w}^0 + \alpha_n\boldsymbol{u}) - \mathrm{RCV}_n(\boldsymbol{w}^0) \\
=&\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M}(w_m^0 + \alpha_n u_m)\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}\right) - \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m^0\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}\right) \\
=&\left\{\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M}(w_m^0 + \alpha_n u_m)\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}\right)\right. \\
&\left. - \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M}(w_m^0 + \alpha_n u_m)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right)\right\} \\
&+ \left\{\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m^0\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right) - \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m^0\boldsymbol{x}_{i(m)}^T\widehat{\Theta}_{(m)}^{[-i]}\right)\right\} \\
&+ \left\{\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M}(w_m^0 + \alpha_n u_m)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right) - \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m^0\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right)\right\} \\
\equiv&\Pi_{n,1} + \Pi_{n,2} + \Pi_{n,3}. \qquad \text{(S4.39)}
\end{aligned}$$

Let

$$L_{in} = h(\breve{\boldsymbol{x}}_i)[\rho\left\{\varepsilon_i^*\left(\boldsymbol{w}^0\right) - \alpha_n \boldsymbol{F}(\boldsymbol{x}_i, \Theta^*)^T \boldsymbol{u}\right\} - \rho\left\{\varepsilon_i^*\left(\boldsymbol{w}^0\right)\right\} + \alpha_n \psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}\boldsymbol{F}(\boldsymbol{x}_i, \Theta^*)^T \boldsymbol{u}],$$

$U_n = \sum_{i=1}^n \mathrm{E}(L_{in})$ and $P_n = \sum_{i=1}^n h(\breve{\boldsymbol{x}}_i)\psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}\boldsymbol{F}(\boldsymbol{x}_i, \Theta^*)$. Thus, $\Pi_{n,3}$ can be decomposed as

$$\begin{aligned}\Pi_{n,3} &= U_n - \alpha_n P_n^T \boldsymbol{u} + \sum_{i=1}^n \{L_{in} - \mathrm{E}(L_{in})\} \\ &\equiv \Lambda_{n,1} + \Lambda_{n,2} + \Lambda_{n,3}. \end{aligned} \tag{S4.40}$$

Combining (S4.39) and (S4.40), we obtain

$$\mathrm{RCV}_n(\boldsymbol{w}^0 + \alpha_n \boldsymbol{u}) - \mathrm{RCV}_n(\boldsymbol{w}^0) = \Pi_{n,1} + \Pi_{n,2} + \Lambda_{n,1} + \Lambda_{n,2} + \Lambda_{n,3}. \tag{S4.41}$$

In the following, we consider each term of (S4.41) and show that $\Lambda_{n,1}$ is its leading term.

We first consider $\Pi_{n,1}$. Let $a_i(\boldsymbol{w}^0) = \mu_i - \sum_{m=1}^M (w_m^0 + \alpha_n u_m)\boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*$.

According to (S3.27), we expand $\Pi_{n,1}$ to the following form

$$\begin{aligned}\Pi_{n,1} =& \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho \left\{ \varepsilon_i + a_i(\boldsymbol{w}^0) + \sum_{m=1}^{M} (w_m^0 + \alpha_n u_m) \boldsymbol{x}_{i(m)}^T \left(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]} \right) \right\} \\ & - \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\rho \left\{ \varepsilon_i + a_i(\boldsymbol{w}^0) \right\} \\ =& \sum_{i=1}^{n} \sum_{m=1}^{M} (w_m^0 + \alpha_n u_m) h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)}^T \left(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]} \right) \psi \left\{ \varepsilon_i + a_i(\boldsymbol{w}^0) \right\} \\ & + \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\sum_{m=1}^{M} (w_m^0 + \alpha_n u_m) \boldsymbol{x}_{i(m)}^T \left(\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]} \right)} \left[\psi \left\{ \varepsilon_i + a_i(\boldsymbol{w}^0) + t \right\} - \psi \left\{ \varepsilon_i + a_i(\boldsymbol{w}^0) \right\} \right] dt \\ \equiv& \Pi_{n,1(1)} + \Pi_{n,1(2)}.\end{aligned}$$

From Assumptions 1 and 5, Cauchy-Schwarz inequality, and Lemma 1, it is seen that

$$\begin{aligned}|\Pi_{n,1(1)}| &\le C \sum_{i=1}^{n} \sum_{m=1}^{M} (w_m^0 + \alpha_n u_m) \| h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)} \| \| \Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]} \| \\ &\le Cn \max_{1 \le i \le n} \max_{1 \le m \le M} \| \Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]} \| \| h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)} \| \\ &= O_p(n^{1/2} M^{1/2}).\end{aligned}$$

Similarly, we can prove that $\Pi_{n,1(2)} = O_p(n^{1/2}M^{1/2})$. Thus, it follows that

$$\Pi_{n,1} = O_p(n^{1/2} M^{1/2}). \tag{S4.42}$$

Similar to the treatment on $\Pi_{n,1}$, we can show that

$$\Pi_{n,2} = O_p(n^{1/2} M^{1/2}). \tag{S4.43}$$

We now turn our attention to $\Lambda_{n,1}$. Some elementary calculations yield that

$$L_{in} = h(\breve{\boldsymbol{x}}_i) \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left[\psi\{\varepsilon_i^*(\boldsymbol{w}^0) + t\} - \psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}\right] dt.$$

By Assumptions 5 and 7, and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\max_{1\le i\le n} |\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}| &\le \alpha_n \|\boldsymbol{u}\| \max_{1\le i\le n} \left\{ \sum_{m=1}^{M} (\boldsymbol{x}_{i(m)}^T \Theta_{(m)}^*)^2 \right\}^{1/2} \\
&\le \alpha_n \|\boldsymbol{u}\| \max_{1\le i\le n} \max_{1\le m\le M} \|\boldsymbol{x}_{i(m)}\| \max_{1\le m\le M} \|\Theta_{(m)}^*\| M^{1/2} \\
&= o(1) \qquad \text{(S4.44)}
\end{aligned}$$

almost surely. Similar to the derivation of (S2.9), we have

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{E}(L_{in}|X) &= \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left(\mathrm{E}[\psi\{\varepsilon_i^*(\boldsymbol{w}^0) + t\}|X] - \mathrm{E}[\psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}|X]\right) dt \\
&= \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left[R_i\left\{\xi_i^*(\boldsymbol{w}^0) + t\right\} - R_i\left\{\xi_i^*(\boldsymbol{w}^0)\right\}\right] dt \\
&= \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} R_{i1}\left\{\xi_i^*(\boldsymbol{w}^0) + \zeta_i(t)\right\} t dt, \qquad \text{(S4.45)}
\end{aligned}$$

where $\zeta_i(t)$ lies between 0 and $-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}$. Thus, by Assumptions 3 and 8, and (S4.44), we obtain

$$\sum_{i=1}^{n} \mathrm{E}(L_{in}|X) \ge \frac{c}{2} \alpha_n^2 \boldsymbol{u}^T \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*) \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u} \ge \frac{c}{2} n \alpha_n^2 \|\boldsymbol{u}\|^2$$

almost surely. This indicates that

$$\Lambda_{n,1} \ge \frac{c}{2} n \alpha_n^2 \|\boldsymbol{u}\|^2 \qquad \text{(S4.46)}$$

We consider $\Lambda_{n,2}$ below. Applying Lagrange Multiplier Method, we find that optimizing $\overline{\mathrm{FPE}}_n(\boldsymbol{w})$ with a constraint $\sum_{m=1}^{M} w_m = 1$ is equivalent to optimizing

$$L(\boldsymbol{w}, \lambda) = \sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i)\rho\left(y_i - \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* \right) \right\} + \lambda \left(\sum_{m=1}^{M} w_m - 1 \right),$$

where λ is a Lagrange multiplier. Since $\boldsymbol{w}^0$ is the interior point of $\mathcal{W}$, there exists a constant λ^0 such that

$$\begin{cases} -\sum\limits_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\} \boldsymbol{x}_{i(1)}^T \Theta_{(1)}^*] + \lambda^0 = 0, \\ -\sum\limits_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\} \boldsymbol{x}_{i(2)}^T \Theta_{(2)}^*] + \lambda^0 = 0, \\ \qquad\qquad \vdots \\ -\sum\limits_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\} \boldsymbol{x}_{i(M)}^T \Theta_{(M)}^*] + \lambda^0 = 0, \\ \sum\limits_{m=1}^{M} w_m^0 - 1 = 0. \end{cases} \tag{S4.47}$$

Hence, (S4.47) implies that

$$\sum_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\} \boldsymbol{x}_{i(1)}^T \Theta_{(1)}^*] = \cdots = \sum_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\} \boldsymbol{x}_{i(M)}^T \Theta_{(M)}^*] = \lambda^0. \tag{S4.48}$$

Since $\sum_{m=1}^{M}(w_m^0 + \epsilon_n u_m) = 1$ and $\sum_{m=1}^{M} w_m^0 = 1$, we conclude that $\sum_{m=1}^{M} u_m =$

0. According to (S4.48), it is seen that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T\boldsymbol{u}] \\
=&\sum_{m=1}^{M} u_m \sum_{i=1}^{n} \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\left\{\varepsilon_i^*(\boldsymbol{w}^0)\right\}\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*] \\
=&\lambda^0\sum_{m=1}^{M} u_m = 0. \qquad \text{(S4.49)}
\end{aligned}$$

If we denote $\zeta_i = \psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}h(\breve{\boldsymbol{x}}_i)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T\boldsymbol{u} - \mathrm{E}[h(\breve{\boldsymbol{x}}_i)\psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T\boldsymbol{u}]$, then it is clear that $\mathrm{E}(\zeta_i) = 0$ and $\zeta_1, ..., \zeta_n$ are mutually independent. So from Assumptions 1 and 8, we have

$$\begin{aligned}
\mathrm{E}\left[\left(\sum_{i=1}^{n}\zeta_i\right)^2 |X\right] &= \sum_{i=1}^{n}\mathrm{Var}\left[\psi\{\varepsilon_i^*(\boldsymbol{w}^0)\}h(\breve{\boldsymbol{x}}_i)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T\boldsymbol{u}|X\right] \\
&\le C\boldsymbol{u}^T\left(\sum_{i=1}^{n} h^2(\breve{\boldsymbol{x}}_i)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)\boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T\right)\boldsymbol{u} \\
&= O(nM)\|\boldsymbol{u}\|^2 \qquad \text{(S4.50)}
\end{aligned}$$

almost surely. As the last equality of (S4.50) is unrelated to X, it is seen that $\mathrm{E}[(\sum_{i=1}^{n}\zeta_i)^2] = O(nM)\|\boldsymbol{u}\|^2$. Hence, it follows that $\sum_{i=1}^{n}\zeta_i = O_p(n^{1/2}M^{1/2})\|\boldsymbol{u}\|$. Combining this and (S4.49), we see that

$$\Lambda_{n,2} = -\alpha_n\sum_{i=1}^{n}\zeta_i = O_p(\alpha_n n^{1/2}M^{1/2})\|\boldsymbol{u}\|. \qquad \text{(S4.51)}$$

We finally consider $\Lambda_{n,3}$. Note that $\mathrm{E}[L_{in} - \mathrm{E}(L_{in})] = 0$ and $L_{in} - \mathrm{E}(L_{in})$ $(i = 1, ..., n)$ are mutually independent. Hence, we can easily verify that $\mathrm{E}[\Lambda_{n,3}^2|X] = \sum_{i=1}^{n}\mathrm{Var}(L_{in}|X)$. From Cauchy-Schwarz inequality, Fubini's the-

orem, Assumptions 2 and 8, and (S4.44), we obtain

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathrm{Var}(L_{in}|X) \\
=&\sum_{i=1}^{n} \mathrm{Var}\left[h(\breve{\boldsymbol{x}}_i) \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left\{ \psi(\varepsilon_i^*(\boldsymbol{w}^0)+t) - \psi(\varepsilon_i^*(\boldsymbol{w}^0)) \right\} dt \Big| X \right] \\
\leq&\sum_{i=1}^{n} h^2(\breve{\boldsymbol{x}}_i) \mathrm{E}\left\{ \left[\int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left\{ \psi(\varepsilon_i^*(\boldsymbol{w}^0)+t) - \psi(\varepsilon_i^*(\boldsymbol{w}^0)) \right\} dt \right]^2 |X \right\} \\
\leq&\sum_{i=1}^{n} h^2(\breve{\boldsymbol{x}}_i) \mathrm{E}\left[|\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}| \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \left\{ \psi(\varepsilon_i^*(\boldsymbol{w}^0)+t) - \psi(\varepsilon_i^*(\boldsymbol{w}^0)) \right\}^2 dt \Big| X \right] \\
=&\sum_{i=1}^{n} h^2(\breve{\boldsymbol{x}}_i) |\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}| \int_0^{-\alpha_n \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}} \mathrm{E}\left[\left\{ \psi[\varepsilon_i^*(\boldsymbol{w}^0)+t] - \psi[\varepsilon_i^*(\boldsymbol{w}^0)] \right\}^2 \Big| X \right] dt \\
=&\sum_{i=1}^{n} \{\alpha_n h(\breve{\boldsymbol{x}}_i) \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \boldsymbol{u}\}^2 o(1) \\
=&o(1)\alpha_n^2 \boldsymbol{u}^T \left(\sum_{i=1}^{n} h^2(\breve{\boldsymbol{x}}_i) \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*) \boldsymbol{F}(\boldsymbol{x}_i,\Theta^*)^T \right) \boldsymbol{u} \\
=&o(n\alpha_n^2 M)\|\boldsymbol{u}\|^2 \qquad \text{(S4.52)}
\end{aligned}
$$

almost surely, where o(1) in the third equation is independent of i because of Assumption 2 and (S4.44). Hence, it follows that $\mathrm{E}[\Lambda_{n,3}^2] = o(n\alpha_n^2 M)\|\boldsymbol{u}\|^2$, which implies that

$$\Lambda_{n,3} = o_p(\alpha_n n^{1/2} M^{1/2})\|\boldsymbol{u}\|. \qquad \text{(S4.53)}$$

Combining (S4.42), (S4.43), (S4.46), (S4.51) and (S4.53), $\Lambda_{n,1}$ asymptotically dominates all of the other terms in (S4.41) when a sufficiently large C_0 is chosen. Since $\Lambda_{n,1}$ is positive, this completes the proof of Theorem 2.

S5. Proof of Corollary 1

Observe that

$$
\begin{aligned}
&\overline{\mathrm{FPE}}_n(\widehat{\boldsymbol{w}}) - \overline{\mathrm{FPE}}_n(\boldsymbol{w}^0) \\
=&\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i)\rho\left(\varepsilon_i + \xi_i^*(\boldsymbol{w}^0) + \sum_{m=1}^{M}(w_m^0 - \widehat{w}_m)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\right) - h(\breve{\boldsymbol{x}}_i)\rho\left(\varepsilon_i + \xi_i^*(\boldsymbol{w}^0)\right)\right\} \\
=&\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{ \sum_{m=1}^{M}(w_m^0 - \widehat{w}_m)h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*\psi(\varepsilon_i + \xi_i^*(\boldsymbol{w}^0))\right\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{ h(\breve{\boldsymbol{x}}_i)\int_0^{\sum_{m=1}^{M}(w_m^0 - \widehat{w}_m)\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*} \left\{\psi(\varepsilon_i + \xi_i^*(\boldsymbol{w}^0) + t) - \psi(\varepsilon_i + \xi_i^*(\boldsymbol{w}^0))\right\} dt\right\} \\
\equiv& D_{n,1} + D_{n,2}.
\end{aligned}
$$

We first consider $D_{n,1}$. By Cauchy-Schwarz inequality, Assumptions 1 and 5, and (3.3), it is seen that

$$
\begin{aligned}
|D_{n,1}| &\le C\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \mathrm{E}\left(|\widehat{w}_m - w_m^0|\|h(\breve{\boldsymbol{x}}_i)\boldsymbol{x}_{i(m)}\|\|\Theta_{(m)}^*\|\right) \\
&\le C\overline{k}^{1/2}\sum_{m=1}^{M} \mathrm{E}\left(|\widehat{w}_m - w_m^0|\right) \\
&\le CM^{1/2}\overline{k}^{1/2}\mathrm{E}\left(\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^0\|\right) \\
&= O(\overline{k}^{1/2}M^{3/4}n^{-1/4}).
\end{aligned}
$$

Similarly, we can prove that

$$
|D_{n,2}| = O(\overline{k}^{1/2}M^{3/4}n^{-1/4}).
$$

This completes the proof of Corollary 1.

S6. Proof of Theorem 3

By the definition of $\boldsymbol{w}^0$, it is clear that $\overline{\mathrm{FPE}}_n(\boldsymbol{w}^0) = \inf_{\boldsymbol{w}\in\mathcal{W}} \overline{\mathrm{FPE}}_n(\boldsymbol{w})$. Similar to (S3.31), we can prove that

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \overline{\mathrm{FPE}}_n(\boldsymbol{w}) \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\rho(\varepsilon_i)\right\},$$

which means that $\overline{\mathrm{FPE}}_n(\boldsymbol{w}^0)$ is bounded away from zero. Thus, by Corollary 1, it is seen that

$$\frac{\overline{\mathrm{FPE}}_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} \overline{\mathrm{FPE}}_n(\boldsymbol{w})} = \frac{\overline{\mathrm{FPE}}_n(\boldsymbol{w}^0) + o(1)}{\overline{\mathrm{FPE}}_n(\boldsymbol{w}^0)} = 1 + o(1).$$

This completes the proof of Theorem 3.

S7. Proof of Theorem 4

We first prove that for any $\delta > 0$, there is a sufficiently large integer N such that for $n > N$,

$$\mathrm{P}\left\{\sup_{\boldsymbol{w}_1\in\widetilde{\mathcal{W}},\boldsymbol{w}_2\in\mathcal{W}\backslash\widetilde{\mathcal{W}}} [\mathrm{RCV}_n(\boldsymbol{w}_1) - \mathrm{RCV}_n(\boldsymbol{w}_2)] < 0\right\} \geq 1-\delta, \tag{S7.54}$$

which implies that with probability approaching to one, $\mathrm{RCV}_n(\boldsymbol{w})$ cannot attain its minimum in the set $\mathcal{W}\backslash\widetilde{\mathcal{W}}$, i.e.,

$$\mathrm{P}(\widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}) \to 1.$$

Denote $\widehat{\varepsilon}_{-i}(\boldsymbol{w}) = y_i - \sum_{m=1}^{M} w_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)}^{[-i]}$. Observe that for any $\boldsymbol{w}_1 \in$

$\widetilde{\mathcal{W}}$ and $\boldsymbol{w}_2 \in \mathcal{W}\backslash\widetilde{\mathcal{W}}$,

$$\begin{aligned}
&\mathrm{RCV}_n(\boldsymbol{w}_1) - \mathrm{RCV}_n(\boldsymbol{w}_2) \\
=&\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\{\widehat{\varepsilon}_{-i}(\boldsymbol{w}_1)\} - \rho\{\widehat{\varepsilon}_{-i}(\boldsymbol{w}_2)\}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\{\widehat{\varepsilon}_{-i}(\boldsymbol{w}_1)\} - \rho\{\varepsilon_i^*(\boldsymbol{w}_1)\}\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\{\varepsilon_i^*(\boldsymbol{w}_2)\} - \rho\{\widehat{\varepsilon}_{-i}(\boldsymbol{w}_2)\}\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\left[\rho\{\varepsilon_i^*(\boldsymbol{w}_1)\} - \rho\{\varepsilon_i^*(\boldsymbol{w}_2)\}\right] \\
\equiv& K_{n,1}(\boldsymbol{w}_1) + K_{n,2}(\boldsymbol{w}_2) + K_{n,3}(\boldsymbol{w}_1, \boldsymbol{w}_2).
\end{aligned}$$

Since $\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}$, we obtain $\sum_{m=1}^{M} w_{1m}\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^* = \mu_i$, $\varepsilon_i^*(\boldsymbol{w}_1) = \varepsilon_i$, and $\varepsilon_i^*(\boldsymbol{w}_1) - \varepsilon_i^*(\boldsymbol{w}_2) = \sum_{m=1}^{M} w_{2m}(\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^* - \mu_i)$ for $i = 1, ..., n$. Denote $L_i(\boldsymbol{w}_2) = h(\breve{\boldsymbol{x}}_i)\{\rho(\varepsilon_i) - \rho(\varepsilon_i + \xi_i^*(\boldsymbol{w}_2)) + \xi_i^*(\boldsymbol{w}_2)\psi(\varepsilon_i)\}$. Then it is seen that

$$\begin{aligned}
K_{n,3}(\boldsymbol{w}_1, \boldsymbol{w}_2) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[L_i(\boldsymbol{w}_2)] - \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\xi_i^*(\boldsymbol{w}_2)\psi(\varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n}\{L_i(\boldsymbol{w}_2) - \mathrm{E}[L_i(\boldsymbol{w}_2)]\} \\
&\equiv H_{n,1}(\boldsymbol{w}_2) + H_{n,2}(\boldsymbol{w}_2) + H_{n,3}(\boldsymbol{w}_2).
\end{aligned}$$

So we have

$$\mathrm{RCV}_n(\boldsymbol{w}_1) - \mathrm{RCV}_n(\boldsymbol{w}_2) = K_{n,1}(\boldsymbol{w}_1) + K_{n,2}(\boldsymbol{w}_2) + H_{n,1}(\boldsymbol{w}_2) + H_{n,2}(\boldsymbol{w}_2) + H_{n,3}(\boldsymbol{w}_2).$$

We prove (S7.54) by showing that

$$\text{(i)} \sup_{\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}} |K_{n,1}(\boldsymbol{w}_1)| = o_p(1);$$

$$\text{(ii)} \sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} |K_{n,2}(\boldsymbol{w}_2)| = o_p(1);$$

$$\text{(iii)} \sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} H_{n,1}(\boldsymbol{w}_2) < -c;$$

$$\text{(iv)} \sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} H_{n,2}(\boldsymbol{w}_2) = o_p(1);$$

$$\text{(v)} \sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} H_{n,3}(\boldsymbol{w}_2) = o_p(1),$$

where c is a positive constant.

We first consider (i). By (S3.27), we see that for $\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}$,

$$\begin{aligned} K_{n,1}(\boldsymbol{w}_1) =& \frac{1}{n} \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \left[\rho \left\{ \varepsilon_i + \sum_{m=1}^{M} w_{1m} \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]}) \right\} - \rho(\varepsilon_i) \right] \\ =& \frac{1}{n} \sum_{i=1}^{n} \psi(\varepsilon_i) \sum_{m=1}^{M} w_{1m} h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]}) \\ & + \frac{1}{n} \sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\sum_{m=1}^{M} w_{1m} \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]})} \{\psi(\varepsilon_i + t) - \psi(\varepsilon_i)\} dt \\ \equiv& K_{n,1(1)}(\boldsymbol{w}_1) + K_{n,1(2)}(\boldsymbol{w}_1). \end{aligned}$$

Thus, by Assumptions 1 and 5, Lemma 1 and Cauchy-Schwarz inequality, we have

$$\begin{aligned} \sup_{\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}} |K_{n,1(1)}(\boldsymbol{w}_1)| &\le C \max_{1 \le i \le n} \max_{1 \le m \le M} \|\hat{\Theta}_{(m)}^{[-i]} - \Theta_{(m)}^*\| \max_{1 \le m \le M} \frac{1}{n} \sum_{i=1}^{n} \|h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)}\| \\ &= O_p(n^{-1/2} M^{1/2}) = o_p(1). \end{aligned}$$

In view of the fact that $|\psi(\varepsilon_i+t)-\psi(\varepsilon_i)| \leq C$, we can show that $\sup_{\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}} |K_{n,1(2)}(\boldsymbol{w}_1)| = o_p(1)$. Consequently, we obtain

$$\sup_{\boldsymbol{w}_1 \in \widetilde{\mathcal{W}}} |K_{n,1}(\boldsymbol{w}_1)| = o_p(1).$$

We next consider (ii). Observe that

$$\begin{aligned} K_{n,2}(\boldsymbol{w}_2) =& \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \left[\rho\{\varepsilon_i^*(\boldsymbol{w}_2)\} - \rho\left\{ \varepsilon_i^*(\boldsymbol{w}_2) + \sum_{m=1}^{M} w_{2m} \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]}) \right\} \right] \\ =& -\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_{2m} h(\breve{\boldsymbol{x}}_i) \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]}) \psi\{\varepsilon_i^*(\boldsymbol{w}_2)\} \\ & - \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\sum_{m=1}^{M} w_{2m} \boldsymbol{x}_{i(m)}^T (\Theta_{(m)}^* - \widehat{\Theta}_{(m)}^{[-i]})} [\psi\{\varepsilon_i^*(\boldsymbol{w}_2)+t\} - \psi\{\varepsilon_i^*(\boldsymbol{w}_2)\}] dt. \end{aligned}$$

Hence, following the argument on $K_{n,1}(\boldsymbol{w}_1)$, we can prove that

$$\sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} |K_{n,2}(\boldsymbol{w}_2)| = o_p(1).$$

We now consider (iii). By Fubini's theorem and Taylor expansion, it is seen that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[L_i(\boldsymbol{w}_2)|X] &= \frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \mathrm{E}\left(\int_0^{\xi_i^*(\boldsymbol{w}_2)} \{\psi(\varepsilon_i) - \psi(\varepsilon_i+t)\} dt | X \right) \\ &= -\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\xi_i^*(\boldsymbol{w}_2)} \mathrm{E}[\psi(\varepsilon_i+t)|X] dt \\ &= -\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\xi_i^*(\boldsymbol{w}_2)} R_i(t) dt \\ &= -\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i) \int_0^{\xi_i^*(\boldsymbol{w}_2)} R_{i1}\{\tau_i(t)\} t dt, \end{aligned}$$

where $\tau_i(t)$ lies between 0 and $\xi_i^*(\boldsymbol{w}_2)$. From Assumptions 10 and 11, we can prove that

$$
\begin{aligned}
&\sup_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}} H_{n,1}(\boldsymbol{w}_2)\\
=&\sup_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}} \mathrm{E}\left[-\frac{1}{n}\sum_{i=1}^{n} h(\breve{\boldsymbol{x}}_i)\int_0^{\xi_i^*(\boldsymbol{w}_2)} R_{i1}\{\tau_i(t)\}tdt\right]\\
\le&-c\frac{1}{n}\inf_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}}\sum_{i=1}^{n}\mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\left[\sum_{m=1}^{M} w_{2m}(\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*-\mu_i)\right]^2\right\}\\
=&-c\frac{1}{n}\inf_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}}\sum_{i=1}^{n}\mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\left[\sum_{m=s_0+1}^{M} w_{2m}(\boldsymbol{x}_{i(m)}^T\Theta_{(m)}^*-\mu_i)\right]^2\right\}\\
=&-c\frac{1}{n}\inf_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}}\sum_{i=1}^{n}\mathrm{E}\left\{h(\breve{\boldsymbol{x}}_i)\left[\breve{\boldsymbol{x}}_i^T\sum_{m=s_0+1}^{M} w_{2m}(\overline{\Theta}_{(m)}-\Theta_0)\right]^2\right\}\\
=&-c\inf_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}}\left(\sum_{m=s_0+1}^{M} w_{2m}(\overline{\Theta}_{(m)}-\Theta_0)\right)^T\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}h(\breve{\boldsymbol{x}}_i)\breve{\boldsymbol{x}}_i\breve{\boldsymbol{x}}_i^T\right\}\left(\sum_{m=s_0+1}^{M} w_{2m}(\overline{\Theta}_{(m)}-\Theta_0)\right)\\
\le&-c\inf_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}}\left\|\sum_{m=s_0+1}^{M} w_{2m}(\overline{\Theta}_{(m)}-\Theta_0)\right\|^2.
\end{aligned}
$$

So, by Assumption 12, we have

$$\sup_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}} H_{n,1}(\boldsymbol{w}_2)\le -c.$$

Similar to the treatment on $CV_{2n}(\boldsymbol{w})$ in the proof of Theorem 1, we can show that

$$\sup_{\boldsymbol{w}_2\in\mathcal{W}\setminus\widetilde{\mathcal{W}}} H_{n,2}(\boldsymbol{w}_2)=o_p(1)$$

and

$$\sup_{\boldsymbol{w}_2 \in \mathcal{W} \setminus \widetilde{\mathcal{W}}} H_{n,3}(\boldsymbol{w}_2) = o_p(1).$$

Combining (i)-(v), we conclude that (S7.54) is correct. This completes the proof of Theorem 4.

S8. Proof of Theorem 5

From Cauchy-Schwarz and Jensen's inequalities, we obtain

$$\begin{aligned}
&n^{-1}\|\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\mu}\|^2 \\
=&n^{-1}\sum_{i=1}^{n}\left(\sum_{m=1}^{M}\widehat{w}_m \boldsymbol{x}_{i(m)}^T \widehat{\Theta}_{(m)} - \sum_{m=1}^{M}\widehat{w}_m \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* + \sum_{m=1}^{M}\widehat{w}_m \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* - \mu_i\right)^2 \\
\leq&2n^{-1}\sum_{i=1}^{n}\sum_{m=1}^{M}\widehat{w}_m \|\boldsymbol{x}_{i(m)}\|^2 \|\widehat{\Theta}_{(m)} - \Theta_{(m)}^*\|^2 + 2n^{-1}\sum_{i=1}^{n}\left(\sum_{m=1}^{M}\widehat{w}_m \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* - \mu_i\right)^2 \\
\equiv&\Delta_{n,1}(\widehat{\boldsymbol{w}}) + \Delta_{n,2}(\widehat{\boldsymbol{w}}).
\end{aligned}$$

For $\Delta_{n,1}(\widehat{\boldsymbol{w}})$, by Lemma 1 and Assumption 6, we see that

$$\begin{aligned}
\Delta_{n,1}(\widehat{\boldsymbol{w}}) &\leq 2 \max_{1\leq m\leq M} \|\widehat{\Theta}_{(m)} - \Theta_{(m)}^*\|^2 \max_{1\leq i\leq n}\max_{1\leq m\leq M} \|\boldsymbol{x}_{i(m)}\|^2 \\
&= o_p(1).
\end{aligned}$$

In view of Theorem 4, it is seen that for any $\delta > 0$, we have

$$\begin{aligned}
\mathrm{P}\left(\Delta_{n,2}(\widehat{\boldsymbol{w}}) < \delta\right) &\geq \mathrm{P}\left(\Delta_{n,2}(\widehat{\boldsymbol{w}}) < \delta, \widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}\right) \\
&= \mathrm{P}\left(2n^{-1}\sum_{i=1}^{n}\left(\sum_{m=1}^{M}\widehat{w}_m \boldsymbol{x}_{i(m)}^T \Theta_{(m)}^* - \mu_i\right)^2 < \delta, \widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}\right) \\
&= \mathrm{P}\left(2n^{-1}\sum_{i=1}^{n}\left(\sum_{m=1}^{s_0}\widehat{w}_m \mu_i - \mu_i\right)^2 < \delta, \widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}\right) \\
&= \mathrm{P}\left(2n^{-1}\sum_{i=1}^{n}\left(\mu_i - \mu_i\right)^2 < \delta, \widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}\right) \\
&= \mathrm{P}\left(\widehat{\boldsymbol{w}} \in \widetilde{\mathcal{W}}\right) \to 1, \qquad \text{(S8.55)}
\end{aligned}$$

which implies that $\Delta_{n,2}(\widehat{\boldsymbol{w}}) = o_p(1)$. Since $\delta > 0$, the last equality of (S8.55) is correct. This completes the proof of Theorem 5.

S9. Proof of Theorem S2

Notice that

$$\left\|\mathrm{IF}\left(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)}\right)\right\| \leq \sum_{m=1}^{M-1}\left|\frac{\partial}{\partial\epsilon} w_{\epsilon,m}^*|_{\epsilon=0}\right|\left(\|\breve{\Theta}_{(m)}\| + \|\breve{\Theta}_{(M)}\|\right) + \sum_{m=1}^{M} w_m^* \|\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})\|.$$

So to prove $\left\|\mathrm{IF}\left(\boldsymbol{z}; \sum_{m=1}^{M} w_m^* \widetilde{\Theta}_{(m)}\right)\right\|$ is bounded, it suffices to show that $|\frac{\partial}{\partial\epsilon} w_{\epsilon,m}^*|_{\epsilon=0}|$ $(m = 1, ..., M-1)$ and $\|\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})\|$ $(m = 1, ..., M)$ are bounded. The boundedness of $\|\mathrm{IF}(\boldsymbol{z}; \breve{\Theta}_{(m)})\|$ can be proved by Assumption 1 and the second part of Assumption 5. We now prove that $|\frac{\partial}{\partial\epsilon} w_{\epsilon,m}^*|_{\epsilon=0}|$

$(m = 1, ..., M-1)$ are bounded. From Assumption S4, we have

$$\|\boldsymbol{B}\| = \left(\sum_{m=1}^{M-1} |\boldsymbol{x}^T (\widetilde{\Theta}_{(m)} - \widetilde{\Theta}_{(M)})|^2 \right)^{1/2} \leq C \tag{S9.56}$$

almost surely. It follows from Assumption 5 that

$$\left| h(\boldsymbol{x}) \boldsymbol{x}^T \sum_{m=1}^{M} w_m^* \mathrm{IF}(\boldsymbol{z}; \widetilde{\Theta}_{(m)}) \right| \leq C. \tag{S9.57}$$

Hence, by Assumption S3, (S9.56) and (S9.57), it is seen that $\|\boldsymbol{J}\| \leq C$. Similarly, we can show that $\|\boldsymbol{K}\|$ and $\|\boldsymbol{L}\|$ are bounded. By Triangle inequality, we have

$$\left\| \frac{\partial}{\partial \epsilon} \boldsymbol{\pi}_\epsilon^* |_{\epsilon=0} \right\| \leq \|\boldsymbol{D}^{-1}\boldsymbol{J}\| + \|\boldsymbol{D}^{-1}\boldsymbol{K}\| + \|\boldsymbol{D}^{-1}\boldsymbol{L}\|.$$

Thus, $\| \frac{\partial}{\partial \epsilon} \boldsymbol{\pi}_\epsilon^* |_{\epsilon=0} \|$ is bounded by Assumption S2, which implies that $| \frac{\partial}{\partial \epsilon} w_{\epsilon,m}^* |$ is bounded for $m = 1, ..., M-1$. This completes the proof of Theorem S2.

S10. Additional simulation studies

S10.1 Simulation A for the case with dependent data

In this subsection, we demonstrate the advantage of the proposed RJMA$_A$ and RJMA$_H$ over other robust model selection and averaging methods for the case of dependent data. Here, the data generating process is based on the same setting as in the main paper, except that the covariates x_{ij}, $j = 2, 3, \cdots, 1000$, are generated from a mixture multidimensional distribution

Table S3: MAEs of various estimators for Simulation A: $\varrho = 0.2$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.003†	2.357	1.256	3.997	**0.956**	2.571	1.669	4.514	2.784	4.555	11.877	9.352	10.943
	0.3	1.138†	2.939	1.934	3.741	**1.054**	3.185	2.326	4.452	3.377	4.342	9.487	9.290	8.959
	0.5	1.264†	2.754	1.473	3.403	**1.202**	2.870	1.794	3.906	2.764	3.913	12.202	9.725	11.327
	0.7	1.416†	3.658	1.781	4.011	**1.355**	3.689	2.449	5.562	3.425	5.156	12.283	10.529	11.554
	0.9	1.944†	3.403	2.330	3.997	**1.868**	3.601	2.547	4.428	3.259	4.492	12.322	10.195	11.499
100	0.1	0.736	1.867	0.705†	2.370	**0.692**	2.089	0.974	3.489	1.925	3.293	11.502	9.180	10.749
	0.3	**0.794**	2.909	1.140	3.199	0.800†	3.080	1.376	4.148	2.284	4.050	12.004	10.073	11.285
	0.5	0.868†	2.553	1.075	2.880	**0.854**	2.869	1.305	3.494	2.054	3.442	11.744	9.356	11.201
	0.7	1.035†	2.547	1.310	2.838	**1.034**	2.770	1.448	3.428	2.065	3.344	10.818	8.629	10.048
	0.9	1.541†	2.790	1.754	3.063	**1.525**	2.961	1.916	3.912	2.370	3.799	12.661	9.995	11.631

$(1-\varrho)\mathcal{N}(0,\Sigma) + \varrho t_1(0,\Sigma)$ with the diagonal elements of Σ being 1, and its off-diagonal elements being $0.5^{|i-j|}$. The error term ε_i has a heavy-tailed distribution $(\sum_{j=2}^{6} x_{ij}^2 + 0.01)^{1/2} \times t(1)$. Table S3 summarizes the simulation results for $\varrho = 0.2$.

From Table S3, we find that RJMA$_H$ is the most favored estimator and followed by RJMA$_A$. The robust Mallows-type model averaging methods MA$_A$ and MA$_H$ rank the third and fourth, respectively. For the other two robust model averaging methods SMA$_A$ and SMA$_H$, their MAEs are larger than those of MS$_A$, MS$_H$ and MC$_p$.

S10.2 Simulation B for the case of nonlinear regression

In this subsection, the true data generating process is assumed to be a non-linear model (Burman and Nolan, 1995)

$$y_i = 2\sin(2\pi x_i) + \zeta_i^2 \varepsilon_i, \quad i = 1, ..., n,$$

where $x_i = i/n$, ζ_i follows the normal distribution $\mathcal{N}(x_i, 1)$ and the term ε_i comes from two heavy-tailed distributions, one is the Student's distribution $t(3)$, and the other is the following distribution, which has mean 0 and infinite variance:

$$G(t) = \begin{cases} \dfrac{1}{2(1-t)^2}, & t \leq 0, \\ 1 - \dfrac{1}{2(1+t)^2}, & t > 0. \end{cases}$$

As in Burman and Nolan (1995), a k^{th} degree polynomial in x is used to fit the data, $k = 0, 1, \cdots, 7$. Then, we construct seven nested candidate models with the regressors $\{1, x\}$, $\{1, x, x^2\}$,..., $\{1, x, x^2, ..., x^7\}$, respectively. Let $n = 50$ and 100. The simulation results are shown in Table S4.

It can be seen from Table S4 that when the error term ε_i comes from the distribution $t(3)$, the most favored estimator is RJMA_A. As for RJMA_H, it usually performs better than MS_H, MA_H, SMA_H, MC_p and HC_p. When the error term ε_i comes from the distribution $G(t)$, we find that RJMA_A remains the most preferred estimator. For the sample size of $n = 50$, the

Table S4: MAEs of various estimators for Simulation B

ε	n	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
t(3)	50	**0.180**	0.202	0.184†	0.241	0.292	0.361	0.310	0.391	0.360	0.361	0.746	0.737	0.744
	100	**0.089**	0.109†	0.109†	0.119	0.211	0.238	0.217	0.244	0.218	0.218	0.601	0.587	0.591
G	50	**0.115**	0.131	0.123†	0.152	0.239	0.266	0.252	0.283	0.225	0.245	0.836	0.835	0.842
	100	**0.050**	0.054	0.052†	0.061	0.162	0.156	0.167	0.147	0.125	0.124	0.651	0.649	0.649

MAEs of RJMA$_H$ are smaller than those of MS$_H$, MA$_H$, SMA$_H$ and HC$_p$. When the sample size increases to $n = 100$, the performance of RJMA$_H$ is superior to that of MA$_H$, MMA, JMA and HRC$_p$. Of all cases considered in Simulation B, the traditional model averaging methods MMA, JMA and HRC$_p$ perform poorly. In summary, in the case where the data generating process is non-linear, the proposed estimators RJMA$_A$ and RJMA$_H$ still have good performance when the error term comes from a heavy-tailed distribution.

S10.3 Simulation results for Cases 2, 4 and 6, and Simulation A with 10% and 30% outlier proportions.

In this subsection, we consider 10% and 30% outlier proportions for Cases 2, 4 and 6, and Simulation A. Tables S5-S12 summarize the simulation results.

From Tables S5-S12, it is seen that the proposed methods RJMA$_A$ and

RJMA_H perform well and often rank in the top two for Cases 2, 4 and 6, and Simulation A.

Table S5: MAEs of various estimators for Case 2: $\varrho = 0.1$

n	R^2	RJMA_A	MS_A	MA_A	SMA_A	RJMA_H	MS_H	MA_H	SMA_H	MC_p	HC_p	MMA	JMA	HRC_p
50	0.1	0.324	0.474	0.321	0.527	**0.260**	0.469	0.311†	0.514	0.480	0.585	0.871	0.641	0.817
	0.3	0.369†	0.597	0.391	0.622	**0.334**	0.593	0.397	0.789	0.528	0.679	1.356	0.966	1.285
	0.5	0.437†	0.771	0.479	1.026	**0.410**	0.766	0.473	1.035	0.601	1.094	2.074	1.714	1.904
	0.7	0.554†	0.934	0.593	1.024	**0.525**	0.939	0.590	1.207	0.720	1.203	4.048	3.480	3.906
	0.9	0.874†	1.526	0.915	1.767	**0.840**	1.551	0.898	2.151	1.132	2.031	7.138	5.592	6.981
100	0.1	0.231	0.336	0.216	0.376	**0.193**	0.360	0.221†	0.341	0.332	0.401	1.089	0.582	0.929
	0.3	0.285†	0.456	0.288	0.474	**0.261**	0.472	0.291	0.487	0.377	0.515	3.066	1.825	2.770
	0.5	0.344†	0.547	0.350	0.554	**0.327**	0.596	0.348	0.589	0.416	0.607	5.042	2.893	4.724
	0.7	0.431†	0.661	0.445	0.710	**0.416**	0.806	0.437	1.080	0.484	1.061	3.226	2.208	2.800
	0.9	0.694†	1.175	0.700	1.215	**0.678**	1.321	0.691	1.401	0.724	1.370	7.555	4.662	6.699

Table S6: MAEs of various estimators for Case 2: $\varrho = 0.3$

n	R^2	RJMA_A	MS_A	MA_A	SMA_A	RJMA_H	MS_H	MA_H	SMA_H	MC_p	HC_p	MMA	JMA	HRC_p
50	0.1	0.295†	0.573	0.349	0.599	**0.269**	0.571	0.369	0.664	0.497	0.692	1.813	1.441	1.688
	0.3	0.388†	0.821	0.493	0.860	**0.377**	0.867	0.549	0.939	0.673	0.913	3.015	2.318	2.783
	0.5	0.488†	0.970	0.607	1.161	**0.477**	1.160	0.679	1.516	0.829	1.360	6.315	4.685	6.091
	0.7	0.615†	1.252	0.784	1.496	**0.611**	1.465	0.867	2.003	1.115	1.861	7.957	5.899	7.216
	0.9	**1.017**	2.638	1.633	2.888	1.019†	2.982	1.752	3.651	2.197	3.302	11.934	9.217	11.132
100	0.1	0.210†	0.339	0.210†	0.450	**0.194**	0.350	0.231	0.466	0.302	0.472	1.796	1.326	1.544
	0.3	**0.279**	0.489	0.302	0.500	0.280†	0.532	0.339	0.660	0.414	0.617	2.592	2.504	2.293
	0.5	**0.364**	0.672	0.399†	0.682	**0.364**	0.718	0.441	0.858	0.514	0.830	4.286	3.315	3.636
	0.7	**0.475**	0.834	0.525	0.863	0.486†	0.977	0.575	1.071	0.668	1.005	6.422	4.393	5.529
	0.9	**0.800**	1.473	0.850	1.527	0.816†	1.776	0.907	2.026	1.094	1.858	8.586	7.812	7.550

Table S7: MAEs of various estimators for Case 4: $\varrho = 0.1$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.439	0.652	0.406†	0.932	**0.391**	0.720	0.529	1.064	1.005	1.173	3.345	3.126	3.282
	0.3	0.485†	0.831	0.493	1.058	**0.448**	0.933	0.604	1.340	1.056	1.343	5.717	4.518	5.320
	0.5	0.582†	0.916	0.593	1.181	**0.544**	1.011	0.698	1.606	1.150	1.520	5.641	4.753	5.228
	0.7	0.722†	1.401	0.772	1.679	**0.683**	1.465	0.889	2.215	1.471	2.075	8.003	6.926	7.461
	0.9	1.057†	1.950	1.175	2.255	**1.047**	2.022	1.251	3.946	1.753	3.200	11.052	10.477	10.322
100	0.1	0.295	0.475	**0.260**	0.595	0.278†	0.614	0.336	0.847	0.662	0.890	3.460	2.947	3.329
	0.3	0.354	0.574	**0.333**	0.715	0.351†	0.729	0.400	0.907	0.680	0.930	5.065	3.972	4.819
	0.5	0.417†	0.726	**0.416**	0.836	0.430	0.901	0.481	1.080	0.769	1.066	5.151	4.458	4.969
	0.7	**0.527**	0.978	0.531†	1.045	0.534	1.168	0.583	1.311	0.848	1.283	7.390	5.618	6.705
	0.9	**0.810**	1.315	0.826	1.387	0.812†	1.567	0.854	1.895	1.071	1.769	10.595	7.101	9.571

Table S8: MAEs of various estimators for Case 4: $\varrho = 0.3$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	0.409†	0.868	0.465	0.978	**0.398**	0.963	0.599	1.253	0.979	1.227	6.282	5.376	6.045
	0.3	**0.502**	1.162	0.600	1.375	0.509†	1.234	0.810	1.673	1.279	1.596	5.533	5.217	5.492
	0.5	0.647†	1.501	0.861	1.702	**0.628**	1.608	1.026	2.208	1.477	2.048	8.445	6.965	7.910
	0.7	**0.797**	1.907	1.014	2.293	0.780†	2.245	1.233	3.011	1.768	2.765	9.482	8.048	9.128
	0.9	1.214†	2.836	1.577	3.207	**1.203**	3.140	1.809	4.458	2.682	4.056	14.887	11.223	13.877
100	0.1	0.279	0.563	**0.269**	0.633	0.278†	0.633	0.349	0.780	0.569	0.772	4.488	3.874	4.198
	0.3	**0.356**	0.778	0.366	0.824	0.361†	0.900	0.466	1.019	0.690	0.975	4.503	3.878	4.223
	0.5	**0.447**	0.939	0.474	0.969	0.455†	1.027	0.573	1.257	0.798	1.184	6.010	5.135	5.700
	0.7	**0.574**	1.079	0.634	1.148	0.588†	1.262	0.738	1.668	1.005	1.507	8.267	6.473	7.465
	0.9	**0.939**	1.926	1.023	2.025	0.964†	2.404	1.144	2.700	1.588	2.429	11.249	9.319	10.461

Table S9: MAEs of various estimators for Case 6: $\varrho = 0.1$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.042	2.523	0.960†	3.900	**0.956**	2.589	1.518	8.902	2.741	7.466	17.333	14.560	17.198
	0.3	1.133	2.760	1.125†	3.443	**1.052**	2.705	1.612	5.145	2.684	4.631	16.287	12.566	15.988
	0.5	1.375†	3.103	1.393	4.015	**1.287**	3.326	1.871	6.174	3.023	5.490	17.068	11.367	16.685
	0.7	1.677†	3.666	1.744	4.192	**1.610**	3.796	2.287	6.503	3.250	5.935	16.605	12.488	16.132
	0.9	2.482†	4.797	2.687	5.210	**2.407**	5.286	3.180	7.197	3.963	6.417	16.224	13.057	15.744
100	0.1	0.681	1.480	**0.555**	2.178	0.634†	1.800	0.835	3.042	1.728	2.945	15.837	8.686	14.859
	0.3	0.808	1.884	**0.714**	2.294	0.775†	2.255	0.968	3.164	1.787	3.024	16.980	10.470	16.137
	0.5	0.978	2.092	**0.875**	2.392	0.951†	2.421	1.180	3.285	1.882	3.091	16.023	10.417	15.339
	0.7	1.213	2.437	**1.135**	2.595	1.166†	2.882	1.390	3.533	2.052	3.352	14.157	9.030	13.270
	0.9	1.882	3.434	1.856†	3.437	**1.829**	3.939	2.130	4.131	2.650	3.954	16.506	10.035	15.731

Table S10: MAEs of various estimators for Case 6: $\varrho = 0.3$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.289†	3.271	1.532	4.345	**1.250**	3.319	2.381	5.908	3.731	5.414	20.290	12.573	17.957
	0.3	1.565†	3.395	1.869	4.144	**1.547**	3.475	2.513	5.811	3.687	5.330	18.730	12.625	16.400
	0.5	1.917†	3.761	2.171	4.579	**1.882**	4.047	2.709	6.168	4.011	5.705	23.978	13.547	21.194
	0.7	2.497†	4.325	2.758	4.976	**2.467**	4.482	3.286	6.816	4.368	6.154	21.350	13.588	18.956
	0.9	4.064†	5.530	4.316	5.722	**4.000**	5.969	4.488	7.195	5.067	6.590	24.054	16.493	22.041
100	0.1	0.985†	2.056	0.980	2.473	**0.956**	2.260	1.452	3.283	2.416	3.157	17.655	9.993	14.082
	0.3	1.235†	2.122	1.246	2.450	**1.201**	2.443	1.622	3.133	2.379	3.010	14.795	9.416	11.657
	0.5	1.506†	2.371	1.566	2.543	**1.485**	2.737	1.798	3.304	2.495	3.163	16.394	10.194	13.363
	0.7	1.965†	2.789	2.028	2.912	**1.925**	3.186	2.191	3.655	2.782	3.474	15.601	9.887	12.220
	0.9	3.126†	3.795	3.136	3.815	**3.032**	4.217	3.160	4.435	3.554	4.237	16.311	11.526	13.940

Table S11: MAEs of various estimators for Simulation A: $\varrho = 0.1$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.005†	1.753	1.044	2.436	**0.877**	1.978	1.542	2.928	2.670	3.168	10.079	8.110	9.107
	0.3	1.042†	1.844	1.129	2.590	**0.932**	2.133	1.527	3.040	2.712	3.246	11.346	9.082	10.239
	0.5	1.134†	1.630	1.152	2.399	**1.033**	1.860	1.446	3.971	2.439	3.597	14.006	11.116	12.691
	0.7	1.336†	2.452	1.394	3.014	**1.248**	2.604	1.673	3.288	2.644	3.539	11.348	9.451	10.270
	0.9	1.795†	2.582	1.950	2.607	**1.776**	2.640	2.129	2.937	2.740	3.038	9.524	8.088	8.409
100	0.1	0.680	2.396	0.655†	2.993	**0.589**	2.657	0.890	4.275	1.799	4.353	10.577	9.016	9.671
	0.3	0.727	1.466	0.706†	1.853	**0.662**	1.702	0.897	2.239	1.705	2.350	10.027	7.612	8.706
	0.5	0.863†	1.569	0.894	2.140	**0.800**	1.823	1.053	2.541	1.766	2.617	10.829	8.639	10.461
	0.7	1.013†	1.651	1.013†	1.841	**0.972**	1.830	1.104	2.270	1.701	2.338	8.642	6.877	7.650
	0.9	1.468†	2.436	1.527	2.503	**1.462**	2.644	1.567	3.057	2.089	3.029	10.310	9.223	9.234

Table S12: MAEs of various estimators for Simulation A: $\varrho = 0.3$

n	R^2	RJMA$_A$	MS$_A$	MA$_A$	SMA$_A$	RJMA$_H$	MS$_H$	MA$_H$	SMA$_H$	MC$_p$	HC$_p$	MMA	JMA	HRC$_p$
50	0.1	1.180†	3.699	1.792	4.696	**1.123**	3.375	2.361	5.558	3.518	5.287	14.255	10.992	13.117
	0.3	1.198†	3.807	1.822	4.586	**1.151**	4.092	2.292	5.774	3.695	5.362	13.206	11.480	12.762
	0.5	1.278†	3.297	2.046	4.477	**1.200**	3.558	2.404	5.188	3.385	4.985	12.022	10.014	11.263
	0.7	1.392†	3.291	2.013	3.773	**1.380**	3.435	2.310	4.747	3.255	4.496	11.670	9.016	10.909
	0.9	1.952†	4.806	2.266	5.176	**1.895**	4.074	2.487	5.826	3.698	5.774	12.911	11.015	12.235
100	0.1	0.734†	4.334	1.530	4.734	**0.710**	4.631	2.284	5.578	3.297	5.399	11.825	10.389	11.420
	0.3	0.814†	2.904	1.084	3.246	**0.788**	3.098	1.423	4.292	2.297	3.962	11.910	8.807	11.483
	0.5	0.890†	2.765	1.100	3.119	**0.868**	3.063	1.304	4.436	2.167	4.229	11.776	9.185	10.882
	0.7	**1.110**	2.821	1.304	3.014	1.131†	3.051	1.502	3.760	2.226	3.628	12.137	9.558	11.253
	0.9	1.611†	3.150	1.696	3.296	**1.584**	3.338	1.758	4.263	2.373	3.960	12.205	9.204	10.896

Bibliography

Burman, P. and D. Nolan (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877–886.

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

Coakley, C. W. and T. P. Hettamansperger (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* **88**, 872–880.

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

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

Feng, Y., Q. F. Liu, Q. S. Yao, and G. Q. Zhao (2021). Model averaging for nonlinear regression models. *Journal of Business and Economic Statistics* **40**, 785–798.

Hampel, F. R. (1968). "Contributions to the theory of robust estimation",

Ph. D. thesis. *University of California Berkeley*. URL: https://books.google.co.uk/books?id=nh_MvwEACAAJ.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

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

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.

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

Mathieu, T. (2022). Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics* **16**, 3695–3750.

Pan, Y. L., H. Y. Wang, K. D. Xu, and H. Huang (2025). Distributed estimation for linear regression with covariates missing at random. *Communications in Statistics - Simulation and Computation* **54**, 583–601.

Simpson, D. G., D. Ruppert, and R. J. Carroll (1992). On one-step GM

estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* **87**, 439–450.

Wang, M. M., K. You, L. X. Zhu, and G. H. Zou (2024). Robust model averaging approach by Mallows-type criterion. *Biometrics* **80**, ujae128.