Online Supplement to "An Efficient Jackknife Model Averaging Method"

Ze Chen, Jun Liao and Wangli Xu

Part A: Discussion on the Potential Exclusion of the Optimal Model for Condition 6

The second part of Condition 6 may preclude the optimal model. For example, assume that the mth model contains the first m regressors and the risk function in terms of the mth model is $m^{-\alpha}n+m$ with $\alpha>1$ (see Cheng et al. (2015)). In this case, the optimal model with the order $n^{1/(1+\alpha)}$ will be precluded.

For this, we next give some discussions and verify that under a restricted weight set, the second part of Condition 6 can be weakened and the optimal model can be included in the candidate model set. Now, similar to Cheng et al. (2015), we consider the following weight set $\mathcal{W}_L=\bigcup_{l=1}^{L}\mathcal{W}_{(l)}$ with $\mathcal{W}_{(l)}=\{w:\tau<w_iI_{\{w_i\neq0\}}\leq1,\sum_{i=1}^{K}I_{\{w_i\neq0\}}=l,\sum_{k=1}^{K}w_k=1\}$ where $0<\tau<1/L$ with L being some fixed integer. Such a weight wet is slightly restricted relative to the general weight set and is considered in the relevant literature such as Cheng et al. (2015). Here, $\mathcal{W}_{(l)}$ can be written as $\mathcal{W}_{(l)}=\bigcup_{1\leq j_1<j_2<\ldots<j_l\leq K}\mathcal{W}_{j_1\ldots j_l}$ with $\mathcal{W}_{j_1\ldots j_l}=\left\{w:w\in\mathcal{W}_{(l)},w_{j_i}\neq0,1\leq i\leq l\right\}$. Also, we consider the nested candidate model set, that is, the kth model uses the first p_k regressors with $1\leq p_1<p_2<\ldots<p_K$. In the case of normal distribution, $KL(w)$ can be equivalently written as $KL(w)=\|\mu-\hat{\mu}(w)\|_2^2$, where $\mu=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}$ and $\hat{\mu}(w)=\sum_{k=1}^{K}w_k\hat{\theta}_{(k)}$. Also, we have $E|\varepsilon_i|^{2G}<C<\infty$ for $G>2L$, where $\varepsilon_i=y_i-\mu_i$. We show that the asymptotic optimality of our model averaging estimator is still valid based on the following conditions.

Condition S1 *$\bar{P}_{(k),ii}\leq C\bar{p}/n$ uniformly in $1\leq k\leq K$ and $1\leq i\leq n$ with $\bar{P}_{(k),ii}$ being the ith diagonal element of $\bar{P}_{(k)}=X_{(k)}(X_{(k)}^{\mathrm{T}}X_{(k)})^{-1}X_{(k)}^{\mathrm{T}}$.*

Condition S2 *$\zeta_n\to\infty$, where $\zeta_n=\min_{1\leq k\leq K}R_k$ with $R_k=E\|\mu-\hat{\mu}_k\|_2^2$, where $\hat{\mu}_k$ is the estimator of μ under the kth candidate model.*

THEOREM S.1. *Suppose that $\bar{p}^2/n=o(1)$ and Conditions S1–S2 hold, then we have $KL(\tilde{w})/\inf_{w\in\mathcal{W}_L}KL(w)\xrightarrow{P}1$ as $n\to\infty$.*

Theorem S.1 shows that the approximate jackknife model averaging estimator is asymptotically optimal in the sense of squared loss over the weight set $\mathcal{W}_L$. For this theorem, the assumption on the risk is $\zeta_n\to\infty$ which is rather mild and the number of covariates is required to have the order $o(n^{1/2})$. Accordingly, the optimal model with the order $n^{1/(1+\alpha)}$ ($\alpha>1$) can be included in the candidate model set.

Proof of Theorem S.1: Let $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)^{\mathrm{T}}$. Recall that $\tilde{\theta}_{(k)}=(\bar{D}_{(k)}(\bar{P}_{(k)}-I_n)+I_n)Y$, where $\bar{P}_{(k)}=X_{(k)}(X_{(k)}^{\mathrm{T}}X_{(k)})^{-1}X_{(k)}^{\mathrm{T}}$, $\bar{D}_{(k)}=\mathrm{diag}\{(1-\bar{P}_{(k),ii})^{-1}\}_{i=1,\ldots,n}$, $\bar{P}_{(k),ii}$ is the ith diagonal element of $\bar{P}_{(k)}$ and $Y=(y_1,\ldots,y_n)^{\mathrm{T}}$. In the case of normal distribution, the weight choice criterion can be equivalently written as

$$\begin{aligned}\arg\max_{w\in\mathcal{W}_L}ACV(w)&=\arg\min_{w\in\mathcal{W}_L}\{\|Y-\tilde{\mu}(w)\|_2^2-\|\varepsilon\|_2^2\}=\arg\min_{w\in\mathcal{W}_L}\{\|\mu-\tilde{\mu}(w)\|_2^2+2\varepsilon^{\top}(\mu-\tilde{\mu}(w))\}\\&=\arg\min_{w\in\mathcal{W}_L}\{\|\mu-\tilde{\mu}(w)\|_2^2+2\varepsilon^{\top}(I_n-\tilde{P}(w))\mu-2\varepsilon^{\top}\tilde{P}(w)\varepsilon\},\end{aligned}$$

where $\tilde{\mu}(w)=\sum_{k=1}^{K}w_k\tilde{\theta}_{(k)}=\tilde{P}(w)Y$, $\tilde{P}(w)=\sum_{k=1}^{K}w_k\tilde{P}_k$ and $\tilde{P}_k=(\bar{D}_{(k)}(\bar{P}_{(k)}-I_n)+I_n)$. Thus, we only need to show that

$$\sup_{w\in\mathcal{W}_L}\left|\varepsilon^{\top}(I_n-\tilde{P}(w))\mu/R(w)\right|=o_p(1),\ \sup_{w\in\mathcal{W}_L}\left|\varepsilon^{\top}\tilde{P}(w)\varepsilon/R(w)\right|=o_p(1),\tag{S.1}$$

$$\sup_{w\in\mathcal{W}_L}\left|\|\mu-\tilde{\mu}(w)\|_2^2/R(w)-1\right|=o_p(1)\text{ and }\sup_{w\in\mathcal{W}_L}\left|\|\mu-\hat{\mu}(w)\|_2^2/R(w)-1\right|=o_p(1).\tag{S.2}$$

Let $\bar{P}(w)=\sum_{k=1}^{K}w_k\bar{P}_{(k)}$. It is seen that

$$\begin{aligned}R(w)&=E\|\mu-\hat{\mu}(w)\|_2^2=\|\mu-\bar{P}(w)\mu\|_2^2+\sigma^2tr(\bar{P}(w)\bar{P}(w))\\&=\sum_{k=1}^{K}\sum_{l=1}^{K}w_kw_l\mu^{\top}(I_n-\bar{P}_{(k)})(I_n-\bar{P}_{(l)})\mu+\sigma^2\sum_{k=1}^{K}\sum_{l=1}^{K}w_kw_ltr(\bar{P}_{(k)}\bar{P}_{(l)})\end{aligned}$$

and $\inf_{w\in\mathcal{W}_{j_1\cdots j_l}} R(w) \geq \tau^2 R_{\max}(j_l)$, where $\sigma^2 = E(\varepsilon_i^2)$ and $R_{\max}(j_l) = \max_{k\in\{j_1,\ldots,j_l\}} R_k$.

First, we focus on the first term of (S.1). Let $Q_{(k)}$ denote the diagonal matrix with the ith diagonal element $\bar{P}_{(k),ii}/(1-\bar{P}_{(k),ii})$, $k = 1, \ldots, K$. Here, $\max_{1\leq k\leq K} \|Q_{(k)}\|_2 = O(\bar{p}/n)$ under Condition S1. Then, by Whittle (1960) and using Condition S2, $\bar{p}^2/n = o(1)$ and $p_k \geq k$ $(k = 1, \ldots, K)$, for any $\delta > 0$, we have

$$
\begin{aligned}
&P\left\{\sup_{w\in\mathcal{W}_L} |\varepsilon^\top(I_n - \tilde{P}(w))\mu/R(w)| \geq \delta\right\} \leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1} P\left\{\sup_{w\in W_{j_1\cdots j_l}} |\varepsilon^\top(I_n - \tilde{P}(w))\mu/R(w)| \geq \delta\right\} \\
&\leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} P\left\{|\varepsilon^\top(I_n - \tilde{P}_k)\mu\{R_{\max}(j_l)\}^{-1}\tau^{-2}| \geq \delta\right\} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} E|\varepsilon^\top(I_n - \tilde{P}_k)\mu|_2^G\{R_{\max}(j_l)\}^{-G} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} \left\{\|(I-\bar{P}_{(k)})\mu\|_2^G\{R_{\max}(j_l)\}^{-G} + \|Q_{(k)}(I-\bar{P}_{(k)})\mu\|_2^G\{R_{\max}(j_l)\}^{-G}\right\} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} (1+\bar{p}^G n^{-G})R_k^{G/2}\{R_{\max}(j_l)\}^{-G} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\{R_{\max}(j_l)\}^{-G/2} \leq C\sum_{l=1}^{L}\sum_{j_l=l}^{\zeta_n}\cdots\sum_{j_1=1}^{j_2-1}\zeta_n^{-G/2} + C\sum_{l=1}^{L}\sum_{j_l=\zeta_n+1}^{K}\cdots\sum_{j_1=1}^{j_2-1} j_l^{-G/2} \\
&\leq C\zeta_n^{-G/2+L} + C\sum_{l=1}^{L}\sum_{j_l=\zeta_n+1}^{K} j_l^{-G/2+l-1} \to 0.
\end{aligned} \tag{S.3}
$$

Thus, the first term of (S.1) is true. Next, we consider the second term of (S.1). By Condition S1, we note that $tr\left(\tilde{P}_k^\top\tilde{P}_k - \bar{P}_{(k)}^\top\bar{P}_{(k)}\right) = tr\left(Q_{(k)}Q_{(k)}(I-\bar{P}_{(k)})\right)$, and thus

$$
\max_{1\leq k\leq K}\left|tr\left(\tilde{P}_k^\top\tilde{P}_k - \bar{P}_{(k)}^\top\bar{P}_{(k)}\right)\right| \leq \max_{1\leq k\leq K}\left\{\|Q_{(k)}\|_2^2(n-p_k)\right\} = O(\bar{p}^2/n). \tag{S.4}
$$

Combining (S.4), Whittle (1960), Condition S2 and the conditions $\bar{p}^2/n = o(1)$ and $p_k \geq k$ $(k = 1, \ldots, K)$, for any $\delta > 0$, we obtain that

$$
\begin{aligned}
&P\left\{\sup_{w\in\mathcal{W}_L} |\varepsilon^\top\tilde{P}(w)\varepsilon/R(w)| \geq \delta\right\} \leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1} P\left\{\sup_{w\in W_{j_1\cdots j_l}} |\varepsilon^\top\tilde{P}(w)\varepsilon/R(w)| \geq \delta\right\} \\
&\leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} P\left\{|\varepsilon^\top\tilde{P}_k\varepsilon\{R_{\max}(j_l)\}^{-1}\tau^{-2}| \geq \delta\right\} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} \left\{tr\left(\tilde{P}_k^\top\tilde{P}_k\right)\right\}^{G/2}\{R_{\max}(j_l)\}^{-G} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} \left[\left\{tr\left(\tilde{P}_k^\top\tilde{P}_k - \bar{P}_{(k)}^\top\bar{P}_{(k)}\right)\right\}^{G/2} + p_k^{G/2}\right]\{R_{\max}(j_l)\}^{-G} \\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} R_k^{G/2}\{R_{\max}(j_l)\}^{-G} + O(\bar{p}^2/n)\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}}\{R_{\max}(j_l)\}^{-G}
\end{aligned}
$$

$$\leq (C+O(\bar{p}^2/n))\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\{R_{\max}(j_l)\}^{-G/2} \leq (C+O(\bar{p}^2/n))\left(\zeta_n^{-G/2+L}+\sum_{l=1}^{L}\sum_{j_l=\zeta_n+1}^{K} j_l^{-G/2+l-1}\right)\to 0,$$

which implies the second term of (S.1).

Now, we consider the first term of (S.2). By $\max_{1\leq k\leq K}\|Q_{(k)}\|_2^{2G} = O(\bar{p}^{2G}/n^{2G})$ and $E\|Y\|_2^{2G} \leq C(\|\mu\|_2^{2G} + E\|e\|_2^{2G}) \leq C(\|\mu\|_2^{2G} + n^G E|e_i|^{2G}) = O(n^G)$, we have $\max_{1\leq k\leq K}\|Q_{(k)}\|_2^{2G}E\|Y\|_2^{2G} = o(1)$ under the condition $\bar{p}^2/n = o(1)$. Then, similar to (S.3), we notice that

$$\begin{aligned}
&P\left\{\sup_{w\in\mathcal{W}_L}\left|\frac{\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2^2}{R(w)}\right|\geq\delta\right\} \leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1} P\left\{\sup_{w\in W_{j_1\cdots j_l}}\left|\frac{\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2^2}{R(w)}\right|\geq\delta\right\}\\
&\leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} P\left\{|\|\tilde{\theta}_{(k)}-\hat{\theta}_{(k)}\|_2^2\{R_{\max}(j_l)\}^{-1}\tau^{-2}|\geq\delta\right\}\\
&= C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}} E\|Q_{(k)}(I-\bar{P}_{(k)})Y\|_2^{2G}\{R_{\max}(j_l)\}^{-G}\\
&\leq \max_{1\leq k\leq K}\|Q_{(k)}\|_2^{2G}E\|Y\|_2^{2G}\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}}\{R_{\max}(j_l)\}^{-G/2}\to 0. \qquad \text{(S.5)}
\end{aligned}$$

Thus, by (S.5) and the fact that

$$\begin{aligned}
&\sup_{w\in\mathcal{W}_L}\left|\|\mu-\tilde{\mu}(w)\|_2^2/R(w)-1\right| \leq \sup_{w\in\mathcal{W}_L}\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2^2/R(w)+2\sup_{w\in\mathcal{W}_L}\|\mu-\hat{\mu}(w)\|_2\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2/R(w)\\
&\leq \sup_{w\in\mathcal{W}_L}\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2^2/R(w)+2\left\{\sup_{w\in\mathcal{W}_L}\|\mu-\hat{\mu}(w)\|_2^2/R(w)\right\}^{1/2}\left\{\sup_{w\in\mathcal{W}_L}\|\tilde{\mu}(w)-\hat{\mu}(w)\|_2^2/R(w)\right\}^{1/2},
\end{aligned}$$

it is seen that the first term of (S.2) is right if the second term of (S.2) holds. Finally, we consider the second term of (S.2). Note that $L(w)-R(w) = \varepsilon^\top\bar{P}(w)\bar{P}(w)\varepsilon - \sigma^2 tr(\bar{P}(w)\bar{P}(w)) - 2\varepsilon^\top\bar{P}(w)(\mu-\bar{P}(w)\mu)$. Further, by Whittle (1960), Condition S2 and $p_k \geq k$ $(k = 1, ..., K)$, for any $\delta > 0$,

$$\begin{aligned}
&P\left\{\sup_{w\in\mathcal{W}_L}\{|\varepsilon^\top\bar{P}(w)\bar{P}(w)\varepsilon-\sigma^2 tr(\bar{P}(w)\bar{P}(w))|/R(w)\}\geq\delta\right\}\\
&\leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1} P\left\{\sup_{w\in W_{j_1\cdots j_l}}\{|\varepsilon^\top\bar{P}(w)\bar{P}(w)\varepsilon-\sigma^2 tr(\bar{P}(w)\bar{P}(w))|/R(w)\}\geq\delta\right\}\\
&\leq \sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k,m\in\{j_1,\ldots,j_l\}} P\left\{|\varepsilon^\top\bar{P}_{(m)}\bar{P}_{(k)}\varepsilon-\sigma^2 tr(\bar{P}_{(m)}\bar{P}_{(k)})|\{R_{\max}(j_l)\}^{-1}\tau^{-2}\geq\delta\right\}\\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k,m\in\{j_1,\ldots,j_l\}}\{tr(\bar{P}_{(m)}\bar{P}_{(m)}\bar{P}_{(k)}\bar{P}_{(k)})\}^{G/2}\{R_{\max}(j_l)\}^{-G}\\
&\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\{R_{\max}(j_l)\}^{-G/2}\to 0,
\end{aligned}$$

and

$$P\left\{\sup_{w\in\mathcal{W}_L}|\varepsilon^\top\bar{P}(w)(I-\bar{P}(w))\mu/R(w)|\geq\delta\right\}\leq\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}P\left\{\sup_{w\in W_{j_1\cdots j_l}}|\varepsilon^\top\bar{P}(w)(I-\bar{P}(w))\mu/R(w)|\geq\delta\right\}$$

$$\leq\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}}\sum_{m\in\{j_1,\ldots,j_l\}}P\left\{|\varepsilon^\top\bar{P}_{(m)}(I-\bar{P}_{(k)})\mu\{R_{\max}(j_l)\}^{-1}\tau^{-2}|\geq\delta\right\}$$

$$\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\sum_{k\in\{j_1,\ldots,j_l\}}\sum_{m\in\{j_1,\ldots,j_l\}}\|\bar{P}_{(m)}(I-\bar{P}_{(k)})\mu\|_2^G\{R_{\max}(j_l)\}^{-G}$$

$$\leq C\sum_{l=1}^{L}\sum_{j_l=l}^{K}\cdots\sum_{j_1=1}^{j_2-1}\{R_{\max}(j_l)\}^{-G/2}\to 0.$$

Thus, the second term of (S.2) is also true. This completes the proof of Theorem S.1.

Additionally, for other exponential family distributions, the asymptotic optimality may be also valid under a weakened version of Condition 6 (similar to Condition S2) in the sense of minimizing the KL loss. However, this problem may be quite tricky, as the specific function $b(x)$ generally takes a more complicated form. Therefore, this warrants further investigation in future research.

Moreover, it is noteworthy that the consistency of the weight estimator (i.e., the weight estimator approaches the optimal weight) is also established (see Theorem 2), for which the second part of Condition 6 is not required. Hence, under this sense of optimality pursued by our model averaging method, the optimal model can be included in the candidate model.

Part B: Justification and Interpretation of Some Conditions

In this section, we have provided some explanations and examples for the third and fourth terms in Condition 1, the second part of Condition 2, Condition 3, Condition 6, Condition 7, and Conditions 10–11.

(1) The third and fourth terms in Condition 1. We first focus on the third term. This condition is also used in Zhang et al. (2016) and Ando and Li (2017). We next give an explanation of rationality under the Gaussian and Bernoulli distributions.

For the Gaussian distribution, $b'(x)=x$. Then using Jensen's inequality, we have

$$\sup_{w\in\mathcal{W}}\sup_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)}n^{-1}\sum_{i=1}^{n}b'\left(\sum_{k=1}^{K}w_kx_{(k),i}^{\mathrm{T}}\beta_{(k)}\right)^2\leq\sup_{w\in\mathcal{W}}\sup_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)}n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}w_k(x_{(k),i}^{\mathrm{T}}\beta_{(k)})^2\triangleq G.$$

Let $\mu^*_{(k)}=(\mu^*_{(k),1},\ldots,\mu^*_{(k),n})^{\mathrm{T}}$ with $\mu^*_{(k),i}=x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}$. When $\delta^*=0$, we derive that $G=\sup_{w\in\mathcal{W}}\sum_{k=1}^{K}w_kn^{-1}\|\mu^*_{(k)}\|^2\leq\max_{1\leq k\leq K}n^{-1}\|\mu^*_{(k)}\|^2$, where the left-hand side of the inequality is bounded if $\max_{1\leq k\leq K}n^{-1}\|\mu^*_{(k)}\|^2$ is bounded. The latter one is an assumption widely used in the literature (see, e.g., Conditions (C.2) and (C.9) of Zhang et al. (2016) and Condition 6 of Zou et al. (2022)).

For the Bernoulli distribution, $b'(x)=e^x/(1+e^x)\leq 1$ for any $x\in R$. Clearly, the third term in Condition 1 holds in this case. Thus, it is reasonable to assume that the third term in Condition 1 holds.

Second, we consider the fourth term in Condition 1, i.e.,

$$\sup_{w\in\mathcal{W}}\sup_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)}n^{-1}\sum_{i=1}^{n}b''\left(\sum_{k=1}^{K}w_kx_{(k),i}^{\mathrm{T}}\beta_{(k)}\right)^2\leq C<\infty. \tag{S.6}$$

For the Gaussian distribution, $b''(x)=1$. In this case, Equation (S.6) holds obviously. For the Bernoulli distribution, $b''(x)=e^x/(1+e^x)^2\leq 1/4$. Clearly, Equation (S.6) also holds in this scenario. To sum up, the third and fourth terms in Condition 1 are mild and reasonable.

(2) The second part of Condition 2. In fact, this condition essentially requires that the random variable $\varepsilon_{(k),i}$ satisfies the subexponential tail probability. Considering that y_i commonly follows well-known subexponential distributions in GLMs (such as the normal, binomial, and Poisson distributions), this condition is therefore reasonable.

(3) Condition 3. That is, there exists a $\delta^* > 0$ such that

$$\min_{1\le k\le K}\inf_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)}\lambda_{\min}(\mathcal{H}_k(\beta_{(k)}))\ge C>0, \tag{S.7}$$

$$\text{and }\min_{1\le i\le n}\min_{1\le k\le K}\inf_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)}\lambda_{\min}(\mathcal{H}^{-i}_k(\beta_{(k)}))\ge C>0. \tag{S.8}$$

This condition is commonly imposed to show the convergence of $\hat{\beta}_{(k)}$ and $\hat{\beta}^{-i}_{(k)}$. The condition stated in (S.7) resembles Assumption (A2) of Liang and Du (2012), Condition (C.4) of Zhang et al. (2016) and Condition 4 of Zou et al. (2022). The condition stated in (S.8) is the cross-validation version of (S.7), which is similar to Condition S.5 of Zou et al. (2025).

For the Gaussian distribution, $b''(x) = 1$. Then (S.7) and (S.8) can be rewritten as

$$\min_{1\le k\le K}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}\right)\ge C>0,\text{ and }\min_{1\le i\le n}\min_{1\le k\le K}\lambda_{\min}\left(\frac{1}{n-1}\sum_{j\neq i}^{n}x_{(k),j}x^{\mathrm{T}}_{(k),j}\right)\ge C>0. \tag{S.9}$$

If K is fixed, the first term of (S.9) is a standard and common assumption to ensure $n^{-1}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}$ is not ill-conditioned. Let $L_{\min}=\min_{1\le i\le n}\lambda_{\min}\big(\frac{1}{n-1}\sum_{j\neq i}^{n}x_{(k),j}x^{\mathrm{T}}_{(k),j}\big)$. Given that $\bar{p}/n=o(1)$, we have $L_{\min}\ge\frac{n}{n-1}\lambda_{\min}\big(\frac{1}{n}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}\big)-\frac{1}{n-1}\|x_{(k),i}\|_2^2\ge\lambda_{\min}\big(\frac{1}{n}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}\big)+o(1)$, implying that the second term of (S.9) also holds. Thus, the assumptions (S.7) and (S.8) are reasonable.

For the Bernoulli distribution, $b''(x)=e^x/(1+e^x)^2$. For all $k\in\{1,\dots,K\}$ and $\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)$, we assume that $\max_{1\le i\le n}|x^{\mathrm{T}}_{(k),i}\beta_{(k)}|<\infty$. This assumption is quite mild, because it essentially requires that the parameter $\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)$ is not abnormal so that the estimator for $\beta_{(k)}$ is well-behaved. Similar conditions can be also found in Zou et al. (2022) (Condition 6) and Zou et al. (2025) (Condition S.3). Then we have $0<C\le b''(x^{\mathrm{T}}_{(k),i}\beta_{(k)})\le 1/4$ for all $\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)$. Thus we can always find some positive constants c_1, c_2, c_3 and c_4 such that for any vector $v_{(k)}\in R^{p_k}$,

$$c_1v^{\mathrm{T}}_{(k)}\frac{1}{n}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}v_{(k)}\le v^{\mathrm{T}}_{(k)}\mathcal{H}_k(\beta_{(k)})v_{(k)}\le c_2v^{\mathrm{T}}_{(k)}\frac{1}{n}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}v_{(k)}$$
$$\text{and }c_3v^{\mathrm{T}}_{(k)}\frac{1}{n-1}\sum_{j\neq i}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}v_{(k)}\le v^{\mathrm{T}}_{(k)}\mathcal{H}^{-i}_k(\beta_{(k)})v_{(k)}\le c_4v^{\mathrm{T}}_{(k)}\frac{1}{n-1}\sum_{j\neq i}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}v_{(k)}.$$

In this situation, we only need to ensure that the minimum eigenvalues of both $n^{-1}\sum_{i=1}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}$ and $(n-1)^{-1}\sum_{j\neq i}^{n}x_{(k),i}x^{\mathrm{T}}_{(k),i}$ are bounded below. Thus, it is realistic to assume that Condition 3 holds.

(4) Condition 6. The first part $\kappa_n\bar{p}/n\to 0$ is relatively weak and the second part restricts that ξ_n has the order larger than $n^{1/2}K^{1/2}\bar{p}$. Similar conditions to the second part are used in Zhang et al. (2016), Ando and Li (2017) and Zou et al. (2022). This condition requires that all candidate models are misspecified. To better understand this condition, suppose that the k^0th model is correctly specified. Thus, we have $\theta^*_{(k^0),i}=\theta_i$. Then it follows that $\xi_n=\inf_{w\in\mathcal{W}}\sum_{i=1}^{n}E_{y^*}\left(\log\left(f\left(y^*_i|\theta_i,\phi\right)\right)-\log\left(f\left(y^*_i|\theta^*_i(w),\phi\right)\right)\right)=0$, and thus the second part of Condition 6 is violated. Therefore, if one of the candidate models is correctly specified, then the second part of Condition 6 does not hold. To illustrate the rationality of the second part, we provide a specific example.

In this example, we consider the identity link, and the regression model can be written as

$$Y=X\beta+\epsilon, \tag{S.10}$$

where $Y=(y_1,\dots,y_n)^{\mathrm{T}}$ is an n-dimensional response vector, $X=(X_1,\dots,X_p)$ is an $n\times p$ design matrix, $X_j=(x_{1j},\dots,x_{nj})^{\mathrm{T}}$, $\beta=(\beta_1,\dots,\beta_p)^{\mathrm{T}}$ is the coefficient vector of size p, and $\epsilon=(\epsilon_1,\dots,\epsilon_n)^{\mathrm{T}}\sim N(0,\sigma^2I_n)$ with the $n\times n$ identity matrix I_n. Clearly, $E(Y)=\mu=X\beta$. Without loss of generality, the predictors and response are centered. Further, to simplify the description, we assume that the design matrix X is orthonormal,

i.e. $n^{-1}X^{\mathrm{T}}X = I_p$. In addition, we assume that the predictors are ordered from most important to least important, that is, $|\beta_j| \geq |\beta_{j'}|$ for $j < j'$.

Suppose that we have K $(K < p)$ nested candidate models, and the kth candidate model includes the first k predictors in $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Let $\mathcal{I}_k$ be the set that contains the indices of the covariates in the kth candidate model. Under the framework of the linear model, $KL(w)$ can be written as $KL(w) = \sum_{i=1}^{n} \big(\mu_i - \sum_{k=1}^{K} w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)}\big)^2 = \sum_{i=1}^{n} (\mu_i - \hat{\mu}_i(w))^2$, where $x_{(k),i}$ is a k-dimensional vector that includes x_{ij} $(j \in \mathcal{I}_k)$. Let $X_{(k)}$ be the design matrix of the kth candidate model. Then, $\hat{\beta}_{(k)} = (X_{(k)}^{\mathrm{T}} X_{(k)})^{-1} X_{(k)}^{\mathrm{T}} Y = n^{-1} X_{(k)}^{\mathrm{T}} Y = (n^{-1} \sum_{i=1}^{n} x_{i1} y_i, \ldots, n^{-1} \sum_{i=1}^{n} x_{ik} y_i)^{\mathrm{T}} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)^{\mathrm{T}}$, and $E(\hat{\beta}_j) = \beta_j$. Also, the pseudo-true parameter $\beta_{(k)}^*$ is $\beta_{(k)}^* = (\beta_{(k),1}^*, \ldots, \beta_{(k),k}^*)^{\mathrm{T}} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$. Thus, we have $\hat{\mu}_i(w) = \sum_{k=1}^{K} w_k \sum_{j=1}^{k} x_{ij} \hat{\beta}_j = \sum_{j=1}^{K} \bar{w}_j x_{ij} \hat{\beta}_j$, where $\bar{w}_j = \sum_{k=j}^{K} w_k$. Further,

$$
\begin{aligned}
KL^*(w) &= \sum_{i=1}^{n} (\mu_i - \mu_i^*(w))^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} x_{ij}\beta_j - \sum_{j=1}^{K} \bar{w}_j x_{ij} \beta_j \right)^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{K} (\beta_j - \bar{w}_j \beta_j) x_{ij} \right)^2 + \sum_{i=1}^{n} \left(\sum_{j=K+1}^{p} x_{ij}\beta_j \right)^2 \\
&= n \sum_{j=1}^{K} (\beta_j - \bar{w}_j \beta_j)^2 + n \sum_{j=K+1}^{p} \beta_j^2 = n \sum_{j=1}^{K} (\beta_j^2 - 2\bar{w}_j \beta_j^2 + \bar{w}_j^2 \beta_j^2) + n \sum_{j=K+1}^{p} \beta_j^2.
\end{aligned}
$$

Further, the optimal weights $\bar{w}_j^*$ can be written as $\bar{w}_1^* = \bar{w}_2^* = \ldots = \bar{w}_K^* = 1$. Therefore, we have $\xi_n = \sum_{j=K+1}^{p} n\beta_j^2$. Then, Condition 6 holds if $K^{1/2}\bar{p} = o(n^{1/2} \sum_{j=K+1}^{p} \beta_j^2)$ and $K\bar{p} = o(n)$. Further, if there exists a $j \in \{K+1, \ldots, p\}$ such that $\beta_j^2 > C > 0$, then the second part of Condition 6 is satisfied when $K^{1/2}\bar{p} = o(n^{1/2})$.

In the following, focusing on the logit and Poisson regression models, we discuss the second part in the case where all candidate models are misspecified. For the logit regression model, we have $\mu_i = b'(\theta_i)$ with $b(\theta_i) = \log(1 + \exp(\theta_i))$ and $\theta_i = x_i^{\mathrm{T}}\beta$. Further, under the logit model, as shown by Yu et al. (2025), there exists a fixed positive constant L_1 such that $KL^*(w) \geq L_1^{-1} \sum_{i=1}^{n} (\mu_i^*(w) - \mu_i)^2$, where $\mu_i^*(w) = b'(\theta_i^*(w))$ with $\theta_i^*(w) = x_i^{\mathrm{T}}\beta^*(w)$ and $\beta^*(w) = \sum_{k=1}^{K} w_k \Pi_{(k)} \beta_{(k)}^*$. By the Taylor expansion at θ_i, we have $KL^*(w) \geq L_1^{-1} \sum_{i=1}^{n} [b''(\bar{\theta}_i^*(w)) x_i^{\mathrm{T}}(\beta^*(w) - \beta)]^2$, where $\bar{\theta}_i^*(w)$ lies between θ_i and $\theta_i^*(w)$. Note that $b''(\bar{\theta}_i^*(w)) = \exp(\bar{\theta}_i^*(w))/(1 + \exp(\bar{\theta}_i^*(w)))^2 > 0$. If there exists a positive constant b_1 such that $b''(\bar{\theta}_i^*(w))$ is uniformly bounded below by b_1, then $KL^*(w) \geq b_1 n L_1^{-1} \lambda_{\min}\big(n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\big) \|\beta^*(w) - \beta)\|_2^2 \geq C b_1 L_1^{-1} n \|\beta^*(w) - \beta)\|_2^2$, where the last inequality holds under the assumption $\lambda_{\min}(n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}) \geq C > 0$. Further, suppose that all candidate models omit at least one true variable, and let β_{omit} denote the non-zero coefficient associated with this omitted variable. Then it follows that $\xi_n = \inf_{w \in \mathcal{W}} KL^*(w) \geq C b_1 L_1^{-1} n \beta_{\text{omit}}^2$, which implies that the second part of Condition 6 holds if $n^{-1/2} K^{1/2} \bar{p} / \beta_{\text{omit}}^2 = o(1)$.

We now turn to the Poisson regression model. In this case, we have $\mu_i = b'(\theta_i)$ with $b(\theta_i) = \exp(\theta_i)$. Also, under the Poisson model, Yu et al. (2025) has verified that there exists a fixed positive constant L_2 such that $KL^*(w) \geq L_2^{-1} \sum_{i=1}^{n} (\mu_i^*(w) - \mu_i)^2$. Similarly, by the Taylor expansion at θ_i and the assumption that there exists a positive constant b_2 such that $b''(\bar{\theta}_i^*(w)) = \exp(\bar{\theta}_i^*(w))$ is uniformly bounded below by b_2, we have $KL^*(w) \geq C b_2 L_2^{-1} n \|\beta^*(w) - \beta)\|_2^2$. Further, $\xi_n = \inf_{w \in \mathcal{W}} KL^*(w) \geq C b_2 L_2^{-1} n \beta_{\text{omit}}^2$. Thus, if $n^{-1/2} K^{1/2} \bar{p} / \beta_{\text{omit}}^2 = o(1)$, then the second part of Condition 6 holds.

Finally, we discuss the rationality of the third part. Without loss of generality, suppose that the first K_0 candidate models are misspecified, while the remaining models are correctly specified. Define $\widetilde{\mathcal{W}} = \{\bar{w} \in [0,1]^{K_0} : \sum_{k=1}^{K_0} \bar{w}_k = 1\}$. Assume that the response variable follows the normal distribution, then $\tilde{\xi}_n$ can be rewritten as

$$
\tilde{\xi}_n = \inf_{w \in \widetilde{\mathcal{W}}} (1 - w_{\text{cor}})^{-2} \sum_{i=1}^{n} \left(\mu_i (1 - w_{\text{cor}}) - \sum_{k=1}^{K_0} w_k x_{(k),i}^{\mathrm{T}} \beta_{(k)}^* \right)^2 = \inf_{w \in \widetilde{\mathcal{W}}} \sum_{i=1}^{n} \left(\mu_i - \sum_{k=1}^{K_0} \frac{w_k}{1 - w_{\text{cor}}} x_{(k),i}^{\mathrm{T}} \beta_{(k)}^* \right)^2.
$$

Note that $\sum_{k=1}^{K_0} w_k/(1 - w_{\text{cor}}) = 1$ and $0 \leq w_k/(1 - w_{\text{cor}}) \leq 1$ for each $k \in \{1, \ldots, K_0\}$. Thus, $\tilde{\xi}_n = \inf_{\bar{w} \in \overline{\mathcal{W}}} \sum_{i=1}^{n} \big(\mu_i - \sum_{k=1}^{K_0} \bar{w}_k x_{(k),i}^{\mathrm{T}} \beta_{(k)}^*\big)^2 = \inf_{\bar{w} \in \overline{\mathcal{W}}} \|\mu - \mu^*(\bar{w})\|_2^2$. In this situation, the third part only requires that the squared loss of the best possible averaging of misspecified models has an appropriate divergent rate,

which is the same as required by the second part. A similar set of conditions is discussed in Zhang and Liu (2023). Hence, it is reasonable to assume that the third part holds.

(6) Condition 7. Similar conditions can also be found in the model averaging and model selection literature (Lv and Liu 2014, Zhang et al. 2016). In Lemma 1, we have derived that $\max_{1\le k\le K}\|(K\bar{p})^{-1/2}n^{1/2}(\hat{\beta}_{(k)}-\beta^*_{(k)})\|_2 = O_p(1)$. So assuming that $\max_{1\le k\le K}E(\|(K\bar{p})^{-1/2}n^{1/2}(\hat{\beta}_{(k)}-\beta^*_{(k)})\|_2^2)\le C$ is also reasonable.

In the following, we show that

$$\max_{1\le k\le K}E(\|\hat{\beta}_{(k)}-\beta^*_{(k)}\|_2^2)\le CK\bar{p}n^{-1} \tag{S.11}$$

for the case of normal response. Note that $\hat{\beta}_{(k)}=(X_{(k)}^{\mathrm{T}}X_{(k)})^{-1}X_{(k)}^{\mathrm{T}}Y$ and $\beta^*_{(k)}=(X_{(k)}^{\mathrm{T}}X_{(k)})^{-1}X_{(k)}^{\mathrm{T}}\mu$. Thus, for each k, we have $E(\|\hat{\beta}_{(k)}-\beta^*_{(k)})\|_2^2)=E(\varepsilon^{\mathrm{T}}X_{(k)}(X_{(k)}^{\mathrm{T}}X_{(k)})^{-2}X_{(k)}^{\mathrm{T}}\varepsilon)\le\frac{\sigma^2}{n}p_k(\lambda_{\min}(n^{-1}X_{(k)}^{\mathrm{T}}X_{(k)}))^{-1}\le\frac{C\sigma^2p_k}{n}$, where the last inequality holds due to $\lambda_{\min}(n^{-1}X_{(k)}^{\mathrm{T}}X_{(k)})\ge C$ uniformly for $k\in\{1,\dots,K\}$. Thus,

$$\max_{1\le k\le K}E(\|\hat{\beta}_{(k)}-\beta^*_{(k)}\|_2^2)\le\sum_{k=1}^{K}E(\|\hat{\beta}_{(k)}-\beta^*_{(k)})\|_2^2)\le\sum_{k=1}^{K}\frac{C\sigma^2\bar{p}}{n}=\frac{C\sigma^2K\bar{p}}{n}.$$

This completes the proof of (S.11). Hence Condition 7 is a realistic assumption.

(7) Condition 10. The first part $\kappa_n\bar{p}/n=o(1)$ is quite mild. Here we focus on the second part, that is, $\psi_n^{1/2}/(n^{\tau}K^{1/2}\bar{p})=o(1)$. It provides the relationship between ψ_n, n, K and $\bar{p}$. Similar condition can also be found in Li et al. (2022).

Assume that the response variable follows the normal distribution. Then we have $\psi_n=\min_{w\in\mathcal{W}}E(\|\mu-\hat{\mu}(w)\|_2^2)$. We consider the practice case where the function form on the covariates is unknown. In the context of nonparametric estimation with twice differentiable regression function, as showed by Stone (1982), the optimal convergence rate for the root mean squared error is $n^{-2/(4+d)}$ with d being the number of regressors, then the lowest risk ψ_n is of the order $n^{d/(4+d)}$. This implies that Condition 10 holds if $n^{d/(8+2d)-\tau}K^{-1/2}\bar{p}^{-1}\to 0$. Specifically, for the case where K and p are fixed, if we take $d/(8+2d)<\tau<1/2$, then $\|\tilde{w}-w^0\|_2\xrightarrow{P}0$.

Moreover, under the linear model (S.10) and the corresponding settings of candidate models, then we have

$$E(\|\mu-\hat{\mu}(w)\|_2^2)=\sum_{i=1}^{n}E\bigg(\sum_{j=1}^{K}(\beta_j-\bar{w}_j\hat{\beta}_j)x_{ij}\bigg)^2+\sum_{i=1}^{n}E\bigg(\sum_{j=K+1}^{p}x_{ij}\beta_j\bigg)^2\triangleq E_1+E_2,$$

where $\bar{w}_j=\sum_{k=j}^{K}w_k$. Further, we see that $E_1=\sum_{i=1}^{n}\sum_{j=1}^{K}\sum_{k=1}^{K}E((\beta_j-\bar{w}_j\hat{\beta}_j)(\beta_k-\bar{w}_k\hat{\beta}_k)x_{ij}x_{ik})=n\sum_{j=1}^{K}E(\beta_j-\bar{w}_j\hat{\beta}_j)^2$, and $E_2=\sum_{i=1}^{n}\sum_{j=K+1}^{p}E(x_{ij}^2\beta_j^2)=n\sum_{j=K+1}^{p}\beta_j^2$. Thus, $E(\|\mu-\hat{\mu}(w)\|_2^2)=n\sum_{j=1}^{K}(\beta_j^2-2\bar{w}_j\beta_j^2+\bar{w}_j^2(\beta_j^2+\sigma^2/n))+n\sum_{j=K+1}^{p}\beta_j^2$. Further, the optimal weights $\bar{w}_j^*$ can be written as $\bar{w}_1^*=1, \bar{w}_j^*=\beta_j^2/(\sigma^2/n+\beta_j^2),$ for $j=2,\dots,K$. Therefore, we have

$$\psi_n=E(\|\mu-\hat{\mu}(\bar{w}^*)\|_2^2)=\sigma^2+\sum_{j=2}^{K}\frac{n\beta_j^2\sigma^2}{\sigma^2+n\beta_j^2}+\sum_{j=K+1}^{p}n\beta_j^2.$$

When K and p are fixed, if the nonzero coefficients of covariates omitted by all of the candidate models are of the order $n^{-\alpha/2}$ with $\alpha>0$, then $\psi_n=O(n^{1-\alpha})$ for $0<\alpha<1$ and $\psi_n=O(1)$ for $\alpha\ge 1$. Thus Condition 10 holds when $\tau>(1-\alpha)/2$ with $0<\alpha<1$ and when $\tau>0$ with $\alpha\ge 1$.

(8) Condition 11. The first part is a common assumption to ensure $n^{-1}\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}$ is not ill-conditioned, which is similar to Assumption (A2) of Liang and Du (2012) and Condition 6 of Zou et al. (2022). The second part is used to guarantee the consistency of $\hat{\beta}(\tilde{w})$. We next give some discussions for the second part.

For the normal distribution, $b''(x)=1$. In this situation, the second part of Condition 11 is naturally satisfied. For the Bernoulli distribution, $b''(x)=e^x/(1+e^x)^2\le 1/4$. To derive the lower bound of $b''(x)$, we assume that

$$\max\bigg\{\max_{1\le i\le n}\max_{1\le k\le K}|x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}|,\max_{1\le i\le n}|x_i^{\mathrm{T}}\beta|\bigg\}\le C<\infty. \tag{S.12}$$

This assumption essentially requires that the parameters $\beta^*_{(k)}$ and β are not abnormal so that the estimator $\hat{\beta}(\tilde{w})$ is well-behaved. The similar assumption is also used in Zou et al. (2022). Based on (S.12) and Lemma 1, if $K\bar{p}^2/n = o(1)$, we have

$$\max_{1\le i\le n}|x_i^{\mathrm{T}}\hat{\beta}(\tilde{w})| \le \max_{1\le i\le n}\left|\sum_{k=1}^{K}\tilde{w}_k x_{(k),i}^{\mathrm{T}}(\hat{\beta}_{(k)}-\beta^*_{(k)})\right| + \max_{1\le i\le n}\left|\sum_{k=1}^{K}\tilde{w}_k x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}\right|$$
$$\le \max_{1\le i\le n}\max_{1\le k\le K}\|x_{(k),i}\|_2\|\hat{\beta}_{(k)}-\beta^*_{(k)}\|_2 + \max_{1\le i\le n}\max_{1\le k\le K}|x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}| \le O_p(K^{1/2}p/n^{1/2}) + C = C(1+o_p(1)).$$

Further we can see that for all $t\in(0,1)$, $\max_{1\le i\le n}|tx_i^{\mathrm{T}}\hat{\beta}(\tilde{w}) + (1-t)x_i^{\mathrm{T}}\beta| \le C(1+o_p(1))$. Thus, $b''(x_i^{\mathrm{T}}(t\hat{\beta}(\tilde{w})+(1-t)\beta))$ has a lower bound with probability tending to one uniformly for $i\in\{1,\ldots,n\}$, which shows that the second part of Condition 11 is valid.

Part C: Some Discussions on the ALL Property

In this section, we provide an example to explain the rationality of the ALL property. Further, motivated by this example, we try to provide some guidance on how to satisfy the ALL property as much as possible.

In this example, we consider a linear regression model with three predictors:

$$Y = X\beta+\varepsilon = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon, \tag{S.13}$$

where Y is an n-dimensional vector of response variable, $X=(X_1,X_2,X_3)$ is an $n\times 3$ design matrix, $X_i=(x_{1i},\ldots,x_{ni})^{\mathrm{T}}$, $\beta=(\beta_1,\beta_2,\beta_3)^{\mathrm{T}}$ is the coefficient vector, and $\epsilon=(\epsilon_1,\ldots,\epsilon_n)^{\mathrm{T}}\sim N(0,\sigma^2 I_n)$ with I_n being the $n\times n$ identity matrix. Let $E(Y)=X\beta\triangleq\mu$. To simplify the description, we assume that the design matrix X is orthonormal, i.e. $n^{-1}X^{\mathrm{T}}X=I_3$. Denote by C a generic positive constant. Further, we assume that $\beta_1^2=Cn^{\alpha_1}$, $\beta_2^2<\beta_1^2/(1+Cn^{\alpha_2})$ and $\beta_2^2/\beta_3^2=o(1)$, where $\alpha_1>-1$ and $\alpha_2>-(\alpha_1+1)/2$.

Now we suppose that there are two candidate models M_1 and M_2. The model M_1 contains the first predictor in (S.13) and the model M_2 contains the second predictor in (S.13). Clearly, M_1 and M_2 are misspecified since they both miss the third predictor in (S.13). Further, the corresponding weight vector $w=(w_1,w_2)^T$ belongs to the set $\mathcal{W}=\left\{w\in[0,1]^2:\sum_{k=1}^2 w_k=1\right\}$. Next, we prove that $\mathcal{W}_S$ obtained by a model screening process based on AIC has the ALL property in this example.

We first calculate the AIC values of models M_1 and M_2, and compare the AIC values between different models. Let $RSS_i\triangleq\|Y-X_i\hat{\beta}_i\|_2^2$ be the residual sum of squares of the model M_i $(i=1,2)$, where $\hat{\beta}_i$ is the maximum likelihood estimator of β_i under the candidate model M_i. Define $I_i\triangleq n\log(RSS_i)-n\log n+2$ as the AIC information criterion of model M_i. Further, the size relationship of I_1 and I_2 can be obtained by comparing RSS_1 and RSS_2. By the basic algebraic calculation,

$$RSS_2-RSS_1 = \|Y-X_2\hat{\beta}_2\|_2^2-\|Y-X_1\hat{\beta}_1\|_2^2 = n(\beta_1^2-\beta_2^2)+2\beta_1X_1^{\mathrm{T}}\varepsilon-2\beta_2X_2^{\mathrm{T}}\varepsilon+n^{-1}(X_1^{\mathrm{T}}\varepsilon)^2-n^{-1}(X_2^{\mathrm{T}}\varepsilon)^2.$$

Using the central limit theorem, we know that $X_1^{\mathrm{T}}\varepsilon=O_p(n^{1/2})$ and $X_2^{\mathrm{T}}\varepsilon=O_p(n^{1/2})$. Further, $\beta_1^2-\beta_2^2>Cn^{\alpha_1+\alpha_2}/(1+Cn^{\alpha_2})$ and $(2\beta_1X_1^{\mathrm{T}}\varepsilon-2\beta_2X_2^{\mathrm{T}}\varepsilon+n^{-1}(X_1^{\mathrm{T}}\varepsilon)^2-n^{-1}(X_2^{\mathrm{T}}\varepsilon)^2)/n=O_p(n^{(\alpha_1-1)/2})$. Since $\alpha_2>-(\alpha_1+1)/2$, $Cn^{\alpha_1+\alpha_2}/(n^{(\alpha_1-1)/2}(1+Cn^{\alpha_2}))\to\infty$ as $n\to\infty$. Thus, for sufficiently large n, we have $RSS_2>RSS_1$. By a model screening process based on AIC, we have $\mathcal{W}_S=\left\{w\in[0,1]^2:w_1=1 \text{ and } w_2=0\right\}$. Next, to show that $\mathcal{W}_S$ is ALL, we only need to prove that $(KL_1-\inf_{w\in\mathcal{W}}KL(w))/\xi_n\xrightarrow{P}0$, where $KL(w)$ has been defined in Section 2, KL_1 is the KL distance between the estimator $\hat{\theta}_1=(\hat{\theta}_{11},\ldots,\hat{\theta}_{n1})^{\mathrm{T}}$ and the true parameter $\theta=(\theta_1,\ldots,\theta_n)^{\mathrm{T}}=X\beta$, $\hat{\theta}_{i1}=x_{i1}\hat{\beta}_1$ and x_{i1} is the ith element of X_1. Denote $\hat{\theta}_2=(\hat{\theta}_{12},\ldots,\hat{\theta}_{n2})^{\mathrm{T}}$, where $\hat{\theta}_{i2}=x_{i2}\hat{\beta}_2$ and x_{i2} is the ith element of X_2.

We first calculate the optimal weights $\hat{w}=(\hat{w}_1,\hat{w}_2)^{\mathrm{T}}$ by solving the following nonlinear programming problem:

$$\begin{cases}\min & KL(w)=\sum_{i=1}^n\sigma^{-2}(\theta_i(\theta_i-w_1\hat{\theta}_{i1}-w_2\hat{\theta}_{i2})-(\theta_i^2/2-(w_1\hat{\theta}_{i1}+w_2\hat{\theta}_{i2})^2/2))\\ \text{s.t.} & g_1(w)=-w_1\le 0, g_2(w)=-w_2\le 0, h_1(w)=w_1+w_2-1=0.\end{cases} \tag{S.14}$$

By solving (S.14), we have

$$\hat{w}_1 = \frac{\sum_{i=1}^n (\theta_i - \hat{\theta}_{i2})(\hat{\theta}_{i1} - \hat{\theta}_{i2})}{\sum_{i=1}^n (\hat{\theta}_{i1} - \hat{\theta}_{i2})^2} = \frac{n\beta_1^2 + \beta_1 X_1^{\mathrm{T}}\varepsilon + \beta_2 X_2^{\mathrm{T}}\varepsilon + n^{-1}(X_2^{\mathrm{T}}\varepsilon)^2}{n(\beta_1^2+\beta_2^2) + 2(\beta_1 X_1^{\mathrm{T}}\varepsilon + \beta_2 X_2^{\mathrm{T}}\varepsilon) + n^{-1}((X_2^{\mathrm{T}}\varepsilon)^2 + (X_1^{\mathrm{T}}\varepsilon)^2)},$$
$$\hat{w}_2 = \frac{\sum_{i=1}^n (\hat{\theta}_{i1} - \theta_i)(\hat{\theta}_{i1} - \hat{\theta}_{i2})}{\sum_{i=1}^n (\hat{\theta}_{i1} - \hat{\theta}_{i2})^2} = \frac{n\beta_2^2 + \beta_1 X_1^{\mathrm{T}}\varepsilon + \beta_2 X_2^{\mathrm{T}}\varepsilon + n^{-1}(X_1^{\mathrm{T}}\varepsilon)^2}{n(\beta_1^2+\beta_2^2) + 2(\beta_1 X_1^{\mathrm{T}}\varepsilon + \beta_2 X_2^{\mathrm{T}}\varepsilon) + n^{-1}((X_2^{\mathrm{T}}\varepsilon)^2 + (X_1^{\mathrm{T}}\varepsilon)^2)}.$$

Further, $\inf_{w\in\mathcal{W}} KL(w) = KL(\hat{w})$. Before proving that $(KL_1 - \inf_{w\in\mathcal{W}} KL(w))/\xi_n \xrightarrow{P} 0$, we first calculate

$$KL_1 - \inf_{w\in\mathcal{W}} KL(w) = KL_1 - KL(\hat{w}) = \sum_{i=1}^n \hat{w}_2(\theta_i(\hat{\theta}_{i2} - \hat{\theta}_{i1}) + \hat{w}_2/2(\hat{\theta}_{i1}^2 - \hat{\theta}_{i2}^2) + \hat{w}_1\hat{\theta}_{i1}(\hat{\theta}_{i1} - \hat{\theta}_{i2})).$$

By the basic algebraic calculation, we have $\sum_{i=1}^n \theta_i(\hat{\theta}_{i2} - \hat{\theta}_{i1}) = n(\beta_2^2 - \beta_1^2) + \beta_2 X_2^{\mathrm{T}}\varepsilon - \beta_1 X_1^T\varepsilon$, $\sum_{i=1}^n \hat{\theta}_{i1}(\hat{\theta}_{i1} - \hat{\theta}_{i2}) = n\beta_1^2 + 2\beta_1 X_1^{\mathrm{T}}\varepsilon + n^{-1}(X_1^{\mathrm{T}}\varepsilon)^2$ and $\sum_{i=1}^n \hat{\theta}_{i1}^2 - \hat{\theta}_{i2}^2 = n(\beta_1^2 - \beta_2^2) + 2(\beta_1 X_1^{\mathrm{T}}\varepsilon - \beta_2 X_2^{\mathrm{T}}\varepsilon) + n^{-1}((X_1^{\mathrm{T}}\varepsilon)^2 - (X_2^{\mathrm{T}}\varepsilon)^2)$. Thus, $KL_1 - \inf_{w\in\mathcal{W}} KL(w) = \sum_{i=1}^n \hat{w}_2(\theta_i(\hat{\theta}_{i2} - \hat{\theta}_{i1}) + \hat{w}_2/2(\hat{\theta}_{i1}^2 - \hat{\theta}_{i2}^2) + \hat{w}_1\hat{\theta}_{i1}(\hat{\theta}_{i1} - \hat{\theta}_{i2})) = \hat{w}_2 O_p(n^{1+\alpha_1})$. If we further calculate and simplify $\hat{w}_2$, for sufficiently large n, we can obtain $\hat{w}_2 \le C(\beta_2^2/(\beta_1^2 + \beta_2^2))$, where C is a some positive constant. We further have

$$KL_1 - \inf_{w\in\mathcal{W}} KL(w) = O_p\left(\frac{n^{1+\alpha_1}\beta_2^2}{\beta_1^2+\beta_2^2}\right). \tag{S.15}$$

Next, we need to calculate the order of ξ_n. To obtain the optimal weights $\tilde{w} = (\tilde{w}_1, \tilde{w}_2)^{\mathrm{T}}$ in the sense of $\inf_{w\in\mathcal{W}} KL^*(w)$, we first calculate $KL^*(w) = \sigma^{-2}((n\beta_1^2 + n\beta_2^2 + n\beta_3^2)/2 - w_1 n\beta_1^2 - w_2 n\beta_2^2 + w_1^2 n\beta_1^2/2 + w_2^2 n\beta_2^2/2)$. Further, we can obtain $\tilde{w}$ by solving the following nonlinear programming problem:

$$\begin{cases} \min & KL^*(w) = \sigma^{-2}((n\beta_1^2 + n\beta_2^2 + n\beta_3^2)/2 - w_1 n\beta_1^2 - w_2 n\beta_2^2 + w_1^2 n\beta_1^2/2 + w_2^2 n\beta_2^2/2), \\ \text{s.t.} & g_1(w) = -w_1 \le 0, g_2(w) = -w_2 \le 0, h_1(w) = w_1 + w_2 - 1 = 0. \end{cases} \tag{S.16}$$

By solving (S.16), we have $\bar{w}_1 = \frac{\beta_1^2}{\beta_1^2+\beta_2^2} > 0$ and $\bar{w}_2 = \frac{\beta_2^2}{\beta_1^2+\beta_2^2} > 0$. Further, by the basic algebraic calculation, for sufficiently large n,

$$\xi_n = KL^*(\bar{w}) = \frac{1}{2}\sigma^{-2}(n\beta_3^2 + \Psi_1\Psi_2) = \Theta((n\beta_3^2 + n\beta_2^2)), \tag{S.17}$$

where $\Psi_1 = (1 + \beta_2^2/\beta_1^2)^{-2}$ and $\Psi_2 = n\beta_2^2(1 + \beta_2^2/\beta_1^2)$.

Now, combining (S.15) and (S.17), we have

$$\frac{KL_1 - \inf_{w\in\mathcal{W}} KL(w)}{\xi_n} = O_p\left(\frac{n^{1+\alpha_1}\beta_2^2}{(\beta_1^2+\beta_2^2)(n\beta_3^2 + n\beta_2^2)}\right) = o_p(1).$$

Given all that, we show that ALL property can be obtained under mild conditions.

Based on this example, it can be seen that if the candidate model set after screening includes those constructed from important variables relative to the models that have been excluded, then the ALL property is more likely to be satisfied. Therefore, to enhance the possibility of achieving the ALL property, it is advisable to first identify the important variables during the model screening process and then construct the candidate model set based on these variables.

Part D: Variable Importance Based on Model Averaging

This section develops a stable variable importance measure based on model averaging. To be specific, according to the candidate model set $\mathcal{M} = \{\mathcal{I}_k, k = 1, \ldots, K\}$ and the estimated weight vector $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_K)^{\mathrm{T}}$, similar in spirit to Chen et al. (2023), we calculate the importance of the jth variable X_j as

$$V_j \triangleq V_j(j; \tilde{w}, \mathcal{M}) = \sum_{k=1}^{K} \tilde{w}_k I(j \in \mathcal{I}_k), \tag{S.18}$$

where $I(\cdot)$ is the indicator function. The importance of the jth variable X_j is quantified as the total weight assigned to candidate models that include X_j, satisfying $0 \le V_j \le 1$. The theoretical property of the proposed importance measure depends on the behavior of the weights $\tilde{w}$, which is given as follows.

COROLLARY S.1. *Under the conditions of Theorem 1, if there is at least one correct model in the candidate model set, then we have* $\min_{j \in \mathcal{I}_T} V_j \xrightarrow{P} 1$, *and* $\max_{j \in \tilde{\mathcal{I}}_F} V_j \xrightarrow{P} 0$ *as* $n \to \infty$, *where* $\tilde{\mathcal{I}}_F = \{1, \ldots, p\} \backslash \tilde{\mathcal{I}}_T$ *and* $\tilde{\mathcal{I}}_T = \{j : j \in \mathcal{I}_k \text{ and } k \in \mathcal{I}_{cor}\}$.

Corollary S.1 shows that the importance of each variable in the true model converges to 1 in probability and the variable importance tends to 0 for the variables outside the correct models.

Proof of Corollary S.1. By the definition of the correct model, we note that the true variables are included in all correct candidate models. Then, for each $j \in \mathcal{I}_T$, we have

$$V_j = \sum_{k=1}^{K} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) = \sum_{k \in \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) + \sum_{k \notin \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) = \tilde{w}_{\text{cor}} + \sum_{k \notin \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right).$$

Further, based on Theorem 1, as $n \to \infty$, we have $\tilde{w}_{\text{cor}} \xrightarrow{P} 1$ and $\sum_{k \neq \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) \le \sum_{k \notin \mathcal{I}_{cor}} \tilde{w}_k \xrightarrow{P} 0$. Thus, $\min_{j \in \mathcal{I}_T} V_j \xrightarrow{P} 1$.

Similarly, for each $j \in \tilde{\mathcal{I}}_F$, as $n \to \infty$, we have $V_j = \sum_{k \in \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) + \sum_{k \notin \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) = \sum_{k \notin \mathcal{I}_{\text{cor}}} \tilde{w}_k I\left(j \in \mathcal{I}_k\right) \xrightarrow{P} 0$. This completes the proof of Corollary S.1.

Next, we give some simulations to evaluate the performance of the proposed variable importance method.

Example D1. In this example, we apply the same logistic regression model as in Example 1. Let $\mathcal{I}_1 = \{1, 2, 3, 4, 5, 7\}$, $\mathcal{I}_2 = \{1, 2, 3, 4, 5\}$, $\mathcal{I}_3 = \{2, 3, 4, 5, 6\}$, $\mathcal{I}_4 = \{3, 4, 5, 6, 7\}$, $\mathcal{I}_5 = \{1, 2, 3, 4, 6\}$, $\mathcal{I}_6 = \{2, 3, 4, 5\}$, $\mathcal{I}_7 = \{3, 4, 5, 6\}$, $\mathcal{I}_8 = \{4, 5, 6, 7\}$, $\mathcal{I}_9 = \{1, 2, 3\}$, $\mathcal{I}_{10} = \{3, 4, 5\}$, $\mathcal{I}_{11} = \{5, 6, 7\}$, $\mathcal{I}_{12} = \{1, 2\}$, $\mathcal{I}_{13} = \{3, 4\}$, $\mathcal{I}_{14} = \{5, 7\}$. The 14 candidate models corresponding to these variable index sets $\mathcal{I}_1, \ldots, \mathcal{I}_{14}$ are considered.

Example D2. In this example, we apply the same Poisson regression model as in Example 2. Let $\mathcal{I}_1 = \{1, 2, 3, 5, 6, 7\}$, $\mathcal{I}_2 = \{1, 2, 3, 4, 5\}$, $\mathcal{I}_3 = \{2, 3, 4, 5, 6\}$, $\mathcal{I}_4 = \{3, 4, 5, 6, 7\}$, $\mathcal{I}_5 = \{1, 2, 3, 4, 6\}$, $\mathcal{I}_6 = \{2, 3, 4, 5\}$, $\mathcal{I}_7 = \{3, 4, 5, 6\}$, $\mathcal{I}_8 = \{4, 5, 6, 7\}$, $\mathcal{I}_9 = \{1, 2, 3\}$, $\mathcal{I}_{10} = \{3, 4, 5\}$, $\mathcal{I}_{11} = \{5, 6, 7\}$, $\mathcal{I}_{12} = \{1, 2\}$, $\mathcal{I}_{13} = \{3, 4\}$, $\mathcal{I}_{14} = \{5, 7\}$. We consider the 14 candidate models corresponding to these variable index sets $\mathcal{I}_1, \ldots, \mathcal{I}_{14}$ in this example.

We set the sample size $n \in \{200, 500, 800, 1000, 1500\}$. The mean importance of each variable is shown in Table S1 based on 200 replications for Examples D1 and D2. Under the setting of Example D1, the important variables are $X_1, X_2, X_3, X_4, X_5, X_7$ and the unimportant variables are X_6. The index sets are $\mathcal{I}_T = \{1, 2, 3, 4, 5, 7\}$ and $\tilde{\mathcal{I}}_F = \{6\}$. It can be seen that the proposed method consistently assigns higher importance values to $X_1, X_2, X_3, X_4, X_5, X_7$ than to X_6 over all the cases. Moreover, as the sample size increases from 200 to 1500, the importance values of $X_1, X_2, X_3, X_4, X_5, X_7$ increase and converge to 1. In contrast, the importance value of X_6 decreases and approaches 0. These results demonstrate that our procedure effectively identifies X_6 as an unimportant variable. For Example D2, the important variables are $X_1, X_2, X_3, X_5, X_6, X_7$ and the unimportant variables are X_4. It is observed from Table S1 that the proposed importance measure identifies this structure accurately.

In the following, we provide some discussions on the comparison between our method and model selection methods. First, model selection methods may have rather high uncertainty when the sample size is relatively small or moderate. For instance, employing cross-validation to determine the optimal tuning parameter in the regularization methods can yield dissimilar results across repeated runs of the same program, owing to the randomness in data splitting. In contrast, our method takes advantage of model averaging to mitigate model selection uncertainty by aggregating the information from multiple models, which may produce more reliable results. Also, Yang and Yang (2017) and Zhang et al. (2024) showed that such weighting based

Table S1 The importance of the variables and the associated standard errors (in parenthesis) for Examples D1 (Logit) and D2 (Poisson).

Logit		X_1	X_2	X_3	X_4	X_5	X_6	X_7
$n=200$	Mean	0.625	0.635	0.491	0.523	0.638	0.154	0.344
	SE	(0.014)	(0.013)	(0.015)	(0.015)	(0.020)	(0.014)	(0.017)
$n=500$	Mean	0.814	0.815	0.752	0.776	0.887	0.114	0.536
	SE	(0.004)	(0.004)	(0.008)	(0.006)	(0.009)	(0.008)	(0.017)
$n=800$	Mean	0.873	0.873	0.840	0.847	0.948	0.079	0.670
	SE	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.007)	(0.016)
$n=1000$	Mean	0.891	0.891	0.871	0.876	0.965	0.069	0.740
	SE	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.011)
$n=1500$	Mean	0.930	0.930	0.916	0.919	0.976	0.046	0.829
	SE	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.007)
Poisson		X_1	X_2	X_3	X_4	X_5	X_6	X_7
$n=200$	Mean	0.580	0.585	0.913	0.352	0.789	0.707	0.767
	SE	(0.017)	(0.017)	(0.010)	(0.018)	(0.010)	(0.011)	(0.010)
$n=500$	Mean	0.823	0.824	0.980	0.174	0.905	0.885	0.895
	SE	(0.009)	(0.009)	(0.002)	(0.011)	(0.003)	(0.004)	(0.004)
$n=800$	Mean	0.902	0.903	0.983	0.093	0.944	0.934	0.941
	SE	(0.005)	(0.005)	(0.001)	(0.007)	(0.002)	(0.002)	(0.002)
$n=1000$	Mean	0.916	0.916	0.988	0.080	0.954	0.945	0.951
	SE	(0.004)	(0.004)	(0.001)	(0.006)	(0.001)	(0.001)	(0.002)
$n=1500$	Mean	0.948	0.948	0.990	0.051	0.970	0.967	0.969
	SE	(0.003)	(0.003)	(0.001)	(0.004)	(0.001)	(0.001)	(0.001)

methods for variable selection become more stable. Second, model selection methods only select a single model among possibly many almost equally good models. When the data are not fully informative, as is often the case for moderate or high-dimensional data, it is more likely that multiple models are equally well supported by the data at hand. Thus, away from glorifying a single selected model, the model averaging based variable importance measure can be used to arrive at a more robust and reliable view based on multiple strong alternative models in terms of the variable importance scores (see, e.g., Zhang et al. (2024) for detailed applications).

Part E: Additional Numerical Results

This sections presents some simulation results as well as real-data analyses for the credit card clients dataset.

E.1. Simulation Studies

Example 3. In this example, we apply the same logistic regression model as in Example 1. Let $\mathcal{I}_1=\{1,2,3,4,5,6,7\}$, $\mathcal{I}_2=\{1,2,3,4,5,7\}$, $\mathcal{I}_3=\{1,2,3,4,5\}$, $\mathcal{I}_4=\{2,3,4,5,6\}$, $\mathcal{I}_5=\{3,4,5,6,7\}$, $\mathcal{I}_6=\{1,2,3,4,6\}$, $\mathcal{I}_7=\{2,3,4,5\}$, $\mathcal{I}_8=\{3,4,5,6\}$, $\mathcal{I}_9=\{4,5,6,7\}$, $\mathcal{I}_{10}=\{1,2,3\}$, $\mathcal{I}_{11}=\{3,4,5\}$, $\mathcal{I}_{12}=\{5,6,7\}$, $\mathcal{I}_{13}=\{1,2\}$, $\mathcal{I}_{14}=\{3,4\}$, $\mathcal{I}_{15}=\{5,7\}$. The 15 candidate models corresponding to these variable index sets $\mathcal{I}_1,\ldots,\mathcal{I}_{15}$ are considered. We set $n\in\{200,500,1000,1500,2000,2500,3000,3500\}$ and $R=200$.

Note that in the setting of Example 3, the sixth attribute with coefficient 0 does not influence the dependent variable, making it an irrelevant variable. Hence, I_1 is regarded as a correct model. The second candidate model is also correctly specified. Other models are misspecified since they omit at least one true variable. Figure S1 presents simulation results on weights and coefficient estimators of EJMA in Example 3. The results in Figure S1 show that the sum of weights on correct models increases with the sample size and approaches one. And the MSE_β value of coefficient estimators decreases and approaches 0 with the increase of the sample size.

Example 4. In this example, we use the same Poisson regression model as in Example 2. Let $\mathcal{I}_1=\{1,2,3,4,5,6,7\}$, $\mathcal{I}_2=\{1,2,3,5,6,7\}$, $\mathcal{I}_3=\{1,2,3,4,5\}$, $\mathcal{I}_4=\{2,3,4,5,6\}$, $\mathcal{I}_5=\{3,4,5,6,7\}$, $\mathcal{I}_6=\{1,2,3,4,6\}$, $\mathcal{I}_7=\{2,3,4,5\}$, $\mathcal{I}_8=\{3,4,5,6\}$, $\mathcal{I}_9=\{4,5,6,7\}$, $\mathcal{I}_{10}=\{1,2,3\}$, $\mathcal{I}_{11}=\{3,4,5\}$, $\mathcal{I}_{12}=\{5,6,7\}$, $\mathcal{I}_{13}=\{1,2\}$, $\mathcal{I}_{14}=\{3,4\}$, $\mathcal{I}_{15}=\{5,7\}$. We consider the 15 candidate models corresponding to these variable index sets $\mathcal{I}_1,\ldots,\mathcal{I}_{15}$ in this example. The sample size is set as $n\in\{100,200,400,600,800,1000,1200,1400\}$ and $R=200$.

Except for the first two models, the other models are misspecified in Example 4. Figure S2 presents the sum of weights on correct models and the MSE_β values of coefficient estimators of EJMA in Example 4. As shown in Figure S2, as n increases, the sum of weights on the first two models gradually converges to

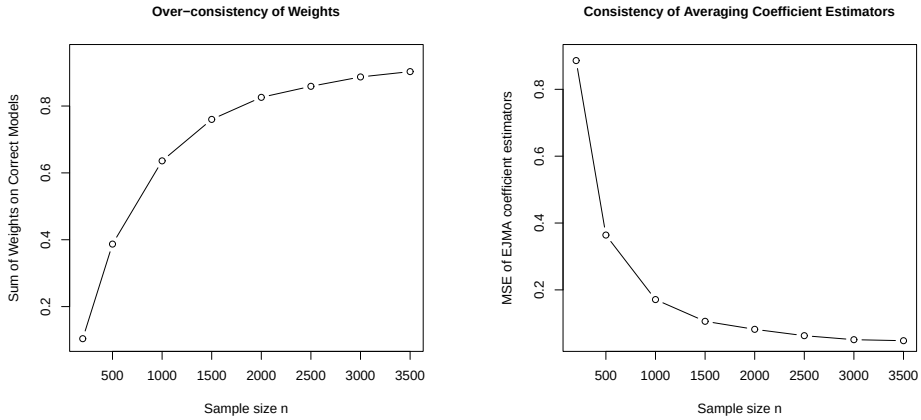

Figure S1 Simulation results on weights and coefficient estimators of EJMA in Example 3.

one and the MSE_β values of coefficient estimators decrease to zero. In summary, the simulation results in Examples 3 and 4 confirm the validity of Theorems 1 and 3.

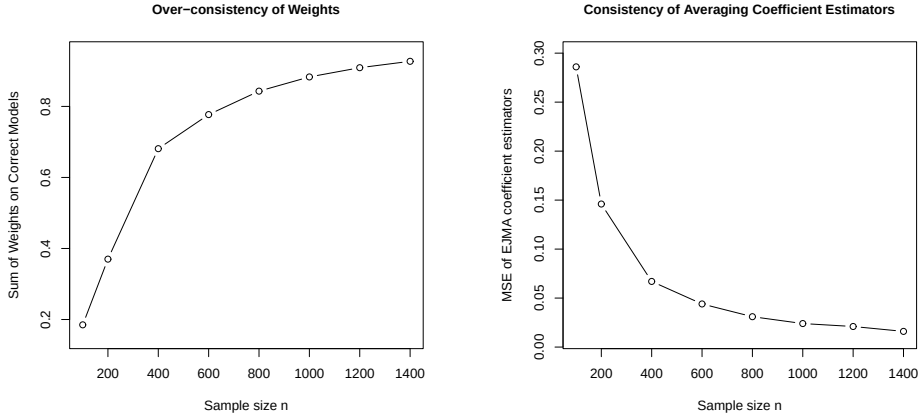

Figure S2 Simulation results on weights and coefficient estimators of EJMA in Example 4.

Example 5. In this example, we consider the logistic regression model: $\mathrm{logit}\{\Pr(y_i = 1)\} = \sum_{j=1}^{p} x_{ij}\beta_j = x_i^{\mathrm{T}}\beta$, where $x_i = (x_{i1}, \ldots, x_{ip})$ are generated from the multivariate normal distribution $N_p(0, \Sigma)$ and $\beta = (1, -0.2, 0.3, 0.5, -0.8, 0, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ with $p = 20$. To mimic the case of model misspecification, the variable x_{i7} is missed when producing the candidate models. With the remaining nine predictors, there are $2^{19} - 1 = 524287$ candidate models, which is computationally infeasible. Thus, the DCMS is applied to prepare the candidate models. The remaining settings are identical to those in Example 1.

Example 6. In this example, we consider the Poisson regression model. The data $\{y_i\}_{i=1}^{n}$ are generated from $\mathrm{Poisson}(\mu_i)$, where $\mu_i = \exp(\sum_{j=1}^{p} x_{ij}\beta_j) = \exp(x_i^{\mathrm{T}}\beta)$, and $\beta = (0.3, 0.15, -0.6, 0, 0.2, -0.05, 0.5, 0, \ldots, 0)^{\mathrm{T}}$. The other settings are the same as in Example 5.

We compare the average computational time of various procedures in Table S2 for Examples 5 and 6. As shown in Table S2, JMA has serious computational burden especially when the sample size is large. In contrast, EJMA has a satisfactory computational efficiency which takes much less computing time relative to JMA. Moreover, EJMA consistently outperforms MACV in terms of computational efficiency across all scenarios. The difference in computational time between them grows with the sample size, indicating that the computational advantage of EJMA becomes more pronounced with larger sample sizes.

Example E1. In this example, we explore the computational advantages of our model averaging method with model screening relative to full subset models. We apply the same logistic and Poisson regression models as in Examples 5 and 6, respectively. Here, we set $n = 200$ and $p = 10$. To mimic the case of model misspecification, the variable x_{i7} is missed when producing the candidate models. With the remaining nine predictors, there are $2^9 - 1 = 511$ candidate models. The remaining settings are same as those in Example 1.

Table S2 The average computational time (in seconds) of various methods for Examples 5 (Logit) and 6 (Poission).

Logit	AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
$n=100$	0.144	0.144	0.143	0.143	0.143	0.435	0.424	3.893	4.848	0.415	0.809
$n=200$	0.309	0.309	0.308	0.308	0.309	0.610	0.630	7.260	9.594	0.619	1.067
$n=300$	0.545	0.545	0.545	0.545	0.545	0.910	0.925	9.697	16.750	0.880	1.406
$n=400$	0.795	0.795	0.795	0.795	0.795	1.173	1.174	10.716	25.263	1.113	1.718
$n=800$	2.181	2.181	2.181	2.181	2.181	2.600	2.589	13.727	80.627	2.600	3.336
Poisson	AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
$n=100$	0.154	0.154	0.154	0.154	0.154	0.431	0.451	1.683	4.397	0.442	0.890
$n=200$	0.314	0.314	0.314	0.314	0.314	0.638	0.658	2.805	10.321	0.620	1.137
$n=300$	0.544	0.544	0.544	0.544	0.544	0.902	0.889	4.053	18.301	0.877	1.484
$n=400$	0.837	0.837	0.837	0.837	0.837	1.221	1.209	5.399	28.326	1.199	1.878
$n=800$	2.648	2.648	2.648	2.648	2.648	3.149	3.190	10.876	90.412	3.145	4.153

Note: The simulations are conducted in R×64 4.1.2 by a computer with an Intel(R) Core(TM) i7-8550U CPU (1.80GHz) with 512 GB memory.

The simulation results in terms of KL loss and computing time for Example E1 are reported in Tables S3 and S4, respectively. Here, Logit_F and Logit_D represent the performance of different methods under the logistic model, based on candidate models obtained from the full subset and DCMS procedures, respectively. Similarly, Poisson_F and Poisson_D denote the corresponding performance under the Poisson model. From Table S3, we can see that with respect to KL loss, EJMA outperforms AIC, BIC, SAIC, SBIC, EMA and OMA_2 by consistently achieving lower values in nearly all cases. Also, Table S3 shows that EJMA, OMA_1, $FJMA_S$, JMA and MACV yield comparable KL loss. However, as shown in Table S4, EJMA exhibits a relatively better computational performance over these methods especially when using the full subset candidate models.

Table S3 The simulation results in terms of KL loss for Example E1.

KL Loss		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
Logit_F	Mean	0.650	0.723	0.558	0.569	0.501	0.508	0.516	0.497	0.494	0.495	0.493
	SE	(0.024)	(0.024)	(0.020)	(0.018)	(0.012)	(0.015)	(0.013)	(0.016)	(0.016)	(0.016)	(0.016)
Logit_D	Mean	0.718	0.735	0.614	0.681	0.571	0.552	0.677	0.551	0.551	0.551	0.553
	SE	(0.020)	(0.023)	(0.019)	(0.019)	(0.015)	(0.019)	(0.024)	(0.018)	(0.018)	(0.018)	(0.018)
Poisson_F	Mean	1.311	1.549	1.235	1.322	1.302	1.220	1.279	1.214	1.214	1.214	1.218
	SE	(0.031)	(0.039)	(0.026)	(0.028)	(0.023)	(0.026)	(0.032)	(0.026)	(0.026)	(0.026)	(0.026)
Poisson_D	Mean	1.411	1.969	1.389	1.880	1.682	1.301	1.706	1.302	1.298	1.299	1.305
	SE	(0.012)	(0.011)	(0.011)	(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)

Note: All values has been magnified by a factor of 10.

Table S4 The simulation results in terms of computing time for Example E1.

Time	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
Logit_F	179.467	182.181	233.498	382.524	147.221	173.815
Logit_D	0.311	0.318	1.917	4.676	0.313	0.504
Poisson_F	167.789	180.826	188.899	429.465	145.505	169.500
Poisson_D	0.335	0.326	0.969	5.507	0.331	0.561

In addition, although EJMA based on the full subset candidate models delivers slightly smaller averaged KL loss values than that with DCMS, the difference is minor. Moreover, EJMA using DCMS demonstrates a significant computational advantage over the full subset version. Thus, we advocate the use of DCMS to screen models prior to model averaging when the number of candidate models is large.

Example E2. In this example, we apply the same logistic regression model as in Example 1 for divergent p. We set $\beta = (1, -0.2, 0.3, 0.5, -0.8, 0, 0.5, 0, 0, \ldots, 0)_p^{\mathrm{T}}$ with $p = [3n^{1/3}]$, where $[x]$ is the smallest integer

larger than x. Here, given that p is relatively large, we adopt the DCMS procedure proposed in Section 4 to prepare candidate models. Other settings are the same as those in Example 1.

Example E3. In this example, we consider the same Poisson regression model as in Example 2 for divergent p. We set $\beta = (0.3, 0.15, -0.6, 0, -0.2, -0.05, 0.5, 0, 0, \ldots, 0)_p^{\mathrm{T}}$ with $p = [3n^{1/3}]$. Also, we use the DCMS procedure to construct candidate models. Other settings are the same as those in Example 2.

The simulation results in terms of KL loss for Examples E2 and E3 are shown in Table S5. From Table S5, it is seen that EJMA always yields smaller KL loss values than their competitors AIC, SAIC, BIC, SBIC and EMA in all cases. The performance gap between OMA_1 and OMA_2 is substantial, highlighting the significant impact of the tuning parameter on the overall performance. In the cases with small sample sizes for the logit model, OMA performs worse than EJMA. $\mathrm{FJMA}_{\mathrm{S}}$ and EJMA yield similar performance in the vast majority of cases. Moreover, as the sample size grows, the MSE_w value of EJMA decreases from 0.00047 to 2e-6 under the logit model, and from 0.00017 to 4e-7 under the Poisson model. This indicates that, for the relatively high-dimensional case, EJMA is still a good approximation to JMA.

Table S5 The simulation results in terms of KL loss for Examples E2 (Logit) and E3 (Poission).

Logit		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$\mathrm{FJMA}_{\mathrm{S}}$	JMA	EJMA	MACV
$n=100$	Mean	14.120	11.075	12.063	10.549	11.014	11.253	10.921	10.641	10.506	10.511	10.444
	SE	(0.967)	(0.394)	(0.808)	(0.363)	(0.376)	(0.481)	(0.384)	(0.428)	(0.406)	(0.415)	(0.376)
$n=200$	Mean	7.893	7.431	7.074	6.966	6.907	6.475	7.094	6.363	6.365	6.365	6.359
	SE	(0.326)	(0.225)	(0.196)	(0.196)	(0.182)	(0.194)	(0.207)	(0.187)	(0.186)	(0.186)	(0.186)
$n=300$	Mean	6.862	6.303	6.147	6.089	6.004	5.414	6.205	5.387	5.385	5.383	5.401
	SE	(0.221)	(0.156)	(0.183)	(0.144)	(0.124)	(0.139)	(0.152)	(0.137)	(0.137)	(0.137)	(0.136)
$n=400$	Mean	6.171	5.592	5.517	5.321	5.320	4.461	5.291	4.449	4.452	4.452	4.460
	SE	(0.188)	(0.115)	(0.166)	(0.107)	(0.102)	(0.108)	(0.108)	(0.104)	(0.104)	(0.104)	(0.106)
$n=800$	Mean	4.975	4.717	4.559	4.624	4.484	3.324	4.594	3.329	3.328	3.327	3.341
	SE	(0.084)	(0.068)	(0.085)	(0.062)	(0.062)	(0.061)	(0.061)	(0.061)	(0.061)	(0.061)	(0.062)
Poisson		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$\mathrm{FJMA}_{\mathrm{S}}$	JMA	EJMA	MACV
$n=100$	Mean	23.451	22.527	20.865	21.749	20.827	19.135	21.785	19.198	19.142	19.145	19.176
	SE	(0.491)	(0.375)	(0.467)	(0.355)	(0.400)	(0.405)	(0.375)	(0.405)	(0.404)	(0.405)	(0.391)
$n=200$	Mean	19.383	20.688	18.041	19.977	18.715	15.323	19.675	15.383	15.355	15.356	15.488
	SE	(0.383)	(0.371)	(0.388)	(0.349)	(0.303)	(0.316)	(0.361)	(0.325)	(0.324)	(0.324)	(0.326)
$n=300$	Mean	15.889	18.806	15.176	18.319	17.025	12.761	17.681	12.792	12.775	12.779	12.903
	SE	(0.279)	(0.296)	(0.270)	(0.276)	(0.253)	(0.206)	(0.307)	(0.231)	(0.228)	(0.229)	(0.233)
$n=400$	Mean	14.237	18.286	13.936	17.960	17.035	12.204	17.259	12.222	12.215	12.217	12.328
	SE	(0.268)	(0.303)	(0.252)	(0.292)	(0.267)	(0.206)	(0.307)	(0.133)	(0.133)	(0.133)	(0.137)
$n=800$	Mean	10.019	16.556	10.019	16.385	15.591	9.414	14.093	9.427	9.426	9.426	9.476
	SE	(0.130)	(0.227)	(0.130)	(0.224)	(0.209)	(0.130)	(0.262)	(0.133)	(0.133)	(0.133)	(0.137)

Note: All values has been magnified by a factor of 100.

In addition, Table S6 reports the average computational time for various methods. From Table S6, we can see that, for the relatively high-dimensional case, our proposed EJMA continues to have a significant improvement over the JMA method in terms of computational cost. Furthermore, EJMA and OMA exhibit the comparable computing efficiency, and they are superior to the other optimal model averaging methods in terms of computational cost.

E.2. A Comparison on Computation Complexity of $\mathrm{FJMA}_{\mathrm{S}}$ and EJMA

We first give an analysis of the computation complexity for $\mathrm{FJMA}_{\mathrm{S}}$. Recall that the log-likelihood function is $F_{(k)}\left(\beta_{(k)}\right) = \sum_{i=1}^{n} \ell\left(y_i|\theta_{(k),i}\right)$ and the estimating equation is $U_{(k)}(\beta_{(k)}) = \partial F_{(k)}(\beta_{(k)})/\partial\beta_{(k)} = 0$. Let $H_{(k)}(\beta_{(k)}) = \partial U_{(k)}(\beta_{(k)})/\partial\beta_{(k)}^{\mathrm{T}}$. The computational cost of $\mathrm{FJMA}_{\mathrm{S}}$ in calculating the approximate jackknife estimator of $\theta_{(k)}$ primarily comes from two aspects, as the costs from other operations are negligible in comparison. The first concerns the calculation of the second-order derivatives of the estimating equation, i.e., $V_{(k)}(\beta_{(k)}) = \{\partial H_{(k)}(\beta_{(k)})/\partial\beta_{(k),1}, \ldots, \partial H_{(k)}(\beta_{(k)})/\partial\beta_{(k),p_k}\}$ with $\beta_{(k),j}$ being the jth element of $\beta_{(k)}$, which requires a computational cost of $O(np_k^3)$. The second is the iterative updating process. It involves the calculation of $V_{(k)}(\hat{\beta}_{(k)})(I_{p_k} \otimes (\hat{\beta}_{(k)} - \beta_{(k)}))$ based on the given $p_k \times p_k^2$ matrix $V_{(k)}(\hat{\beta}_{(k)})$, where $\otimes$

Table S6 The average computational time (in seconds) of various methods for Examples E2 (Logit) and E3 (Poission).

Logit	AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
$n=100$	0.077	0.077	0.077	0.077	0.077	0.193	0.194	1.445	2.201	0.187	0.371
$n=200$	0.193	0.193	0.193	0.193	0.193	0.373	0.382	3.924	6.396	0.375	0.641
$n=300$	0.388	0.388	0.388	0.388	0.388	0.678	0.700	7.857	13.718	0.653	1.037
$n=400$	0.816	0.816	0.816	0.816	0.816	1.303	1.302	15.732	31.237	1.292	1.925
$n=800$	2.989	2.989	2.989	2.989	2.989	3.997	4.008	41.246	132.246	3.977	5.328
Poisson	AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
$n=100$	0.103	0.103	0.103	0.103	0.103	0.254	0.251	0.776	3.159	0.246	0.510
$n=200$	0.268	0.268	0.268	0.268	0.268	0.528	0.542	2.133	10.003	0.517	0.940
$n=300$	0.521	0.521	0.521	0.521	0.521	0.904	0.891	4.348	21.388	0.884	1.475
$n=400$	0.849	0.849	0.849	0.849	0.849	1.335	1.328	7.163	37.384	1.330	2.102
$n=800$	2.803	2.803	2.803	2.803	2.803	3.725	3.789	21.472	142.287	3.725	5.147

Note: The simulations are conducted in R×64 4.1.2 by a computer with an Intel(R) Core(TM) i7-8550U CPU (1.80GHz) with 512 GB memory.

denotes the Kronecker product. This calculation requires a computational cost of $O(p_k^3)$. If this process requires d iterations, then the total computational cost is $O(dp_k^3)$. Thus, the overall computational complexity of FJMA_S in calculating the approximate jackknife estimator of $\theta_{(k)}$ is $O(np_k^3 + dp_k^3)$. As analyzed in Remark 1 of our paper, the computational complexity of our proposed method EJMA in calculating the approximate jackknife estimator of $\theta_{(k)}$ is $O(np_k^2)$. Therefore, from a computational perspective, EJMA is more advantageous in relatively high-dimensional settings.

E.3. Descriptions of Variables in the Vehicle Silhouettes Data

This subsection provides detailed descriptions of variables in the vehicle silhouettes data as follows: elongatedness X_1, skewness about major axis X_2, kurtosis about major axis X_3, kurtosis about minor axis X_4, hollows ratio X_5, scaled radius of gyration X_6, scatter ratio X_7, compactness X_8, pr.axis rectangularity X_9, max.length rectangularity X_{10}, scaled variance along major X_{11}, scaled variance along minor X_{12}, max.length aspect ratio X_{13}, circularity X_{14}, skewness about minor axis X_{15}, pr.axis aspect ratio X_{16}, distance circularity X_{17}, and radius ratio X_{18}.

E.4. Real Data Analysis for the Credit Card Clients Dataset

In this subsection, we apply the proposed EJMA to the credit card clients dataset with 30000 observations to study the behaviors of customers' defaulting on payment in Taiwan. This dataset has been studied by Yeh and Lien (2009) and Zhang and Liu (2023) and can be available at the UC Irvine Machine Learning Repository. The response variable $y_i = 1$ ($y_i = 0$) indicates the defaulting on payment (or not) by a credit card client. There are 23 explanatory variables: amount of the given credit X_1 (NT dollar), gender X_2 (0 = male, 1 = female), education X_3 (1 = university or above, 0 = other), marital status X_4 (1 = married, 0 = other), age X_5 (year), the repayment status in September to April X_6–X_{11}, the amount of bill statement in September to April X_{12}–X_{17}, and the previous amount paid in September to April X_{18}–X_{23}.

Following Yeh and Lien (2009), we use the logistic regression to estimate the probability of defaulting on payment. We randomly use n_{train} observations from the credit card data to estimate the probability of defaulting on payment and model parameters for each candidate model, and then randomly use n_{test} observations as an evaluation set. Note that it is computationally impractical to consider all the subset models as the candidate models for this dataset. Thus, to confront this problem, we construct 23 nested candidate models by using the proposed model screening procedure in Section 4 based on the training data for each repetition. For the assessment of each method in the evaluation set, as in Zhang et al. (2016) and Chen et al. (2023), we also use the following KL-type loss function

$$\text{KL}_{\text{real}} = -2n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}}\log f(y_{\text{test},i}|\hat{\theta}_{\text{test},i}(\hat{w})), \tag{S.19}$$

where $y_{\text{test},i}$ is the ith sample in the evaluation set, $\hat{\theta}_{\text{test},i}(\hat{w}) = \sum_{k=1}^{K}\hat{w}_k x_{\text{test},ki}^{\text{T}}\hat{\beta}_{(k)}$, $x_{\text{test},ki}$ is the ith testing observation for the kth model, and $\hat{w}$ and $\hat{\beta}_{(k)}$ are obtained based on the training set. We repeat the above

steps 200 times and calculate the mean KL-type loss values defined in (S.19) of various methods (i.e., AIC, BIC, SAIC, SBIC, EMA, OMA_1, OMA_2, FJMA_S, JMA, EJMA and MACV).

To better evaluate the performance of each method, we consider various sample sizes for the training set and testing set, that is, $n_{\text{train}} = \{250, 500, 1000, 2000\}$ and $n_{\text{test}} = 5 \times n_{\text{train}} = \{1250, 2500, 5000, 10000\}$. Also, following a referee's suggestion, we consider the larger sample sizes for the training set, that is, $n_{\text{train}} = \{5000, 10000, 15000\}$ and $n_{\text{test}} = 10000$. Table S7 presents the KL-type loss values of various methods and the MSE_w values. It can be observed that JMA produces smaller KL_{real} values than AIC, BIC, SAIC, SBIC and EMA in most cases. Also, the prediction performance of EJMA in terms of KL-type loss is quite similar to that of JMA. For $n = 250, 500, 1000$ and 2000, the MSE_w values are 0.0658, 0.0563, 0.0337 and 0.0208. These relatively small MSE_w values indicate that the weight estimator of EJMA closely approximates that of JMA.

Table S7 The numerical results in terms of KL loss for the credit card clients dataset.

n_{train}		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
250	Mean	10.518	9.874	10.374	9.775	10.406	10.033	9.717	9.928	9.869	9.865	9.841
	SE	(0.098)	(0.038)	(0.084)	(0.032)	(0.053)	(0.054)	(0.029)	(0.055)	(0.042)	(0.041)	(0.039))
500	Mean	9.954	9.582	9.867	9.570	9.733	9.645	9.551	9.627	9.569	9.571	9.575
	SE	(0.035)	(0.016)	(0.033)	(0.015)	(0.024)	(0.026)	(0.019)	(0.025)	(0.019)	(0.020)	(0.020)
1000	Mean	9.628	9.539	9.589	9.525	9.498	9.511	9.490	9.483	9.471	9.470	9.469
	SE	(0.014)	(0.010)	(0.015)	(0.010)	(0.012)	(0.011)	(0.010)	(0.013)	(0.011)	(0.011)	(0.011)
2000	Mean	9.495	9.512	9.478	9.497	9.421	9.425	9.452	9.413	9.413	9.413	9.413
	SE	(0.009)	(0.007)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
5000	Mean	9.363	9.423	9.358	9.414	9.350	9.350	9.391	9.342	—	9.341	9.341
	SE	(0.009)	(0.007)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	—	(0.008)	(0.008)
10000	Mean	9.335	9.374	9.333	9.369	9.342	9.331	9.351	9.330	—	9.328	9.328
	SE	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	—	(0.007)	(0.007)
15000	Mean	9.326	9.335	9.325	9.333	9.335	9.317	9.332	9.317	—	9.317	9.317
	SE	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	—	(0.007)	(0.007)

Note: All values has been magnified by a factor of 10.

Also, we provide the average computational time (in seconds) of various methods in Table S8. It is readily seen that EJMA has a tremendous computational advantage over JMA especially when the sample size of the training set is large. For instance, when $n_{\text{train}} = 2000$, EJMA takes only 2.061 seconds which is about 257 times faster than the 530.371 seconds of JMA. Compared with MACV, it can be seen that the computational benefit of EJMA is evident in all cases. Also, this benefit becomes increasingly significant as the sample size increases.

Table S8 The average computational time (in seconds) of various methods for the credit card clients dataset.

n_{train}	AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
250	0.100	0.100	0.099	0.099	0.099	4.022	4.095	11.322	21.244	0.625	1.449
500	0.142	0.143	0.142	0.142	0.142	7.223	7.053	17.185	55.238	0.825	1.981
1000	0.247	0.247	0.246	0.246	0.246	14.264	14.084	27.032	175.410	1.309	3.202
2000	0.418	0.417	0.416	0.416	0.417	27.172	26.216	39.229	530.371	2.061	5.195
5000	0.888	0.888	0.887	0.887	0.888	69.034	66.688	71.006	—	4.377	11.242
10000	1.647	1.647	1.646	1.647	1.647	146.709	139.196	124.030	—	8.272	20.981
15000	2.095	2.095	2.094	2.094	2.095	201.821	193.501	152.241	—	10.696	26.895

Note: The simulations are conducted in R×64 4.1.2 by a computer with an Intel(R) Core(TM) i7-8550U CPU (1.80GHz) with 512 GB memory.

Part F: Extension to Non-exponential Family Distribution

This section extends our model averaging procedure to the non-exponential family distribution (e.g., the negative binomial distribution). To be specific, for a potentially non-exponential family distribution, a model averaging framework based on CV can be constructed as follows. Suppose we have n independent

and identically distributed (i.i.d.) observations $\{(y_i, x_i)\}_{i=1}^n$. Here, y_i is a scalar response variable, which can be binary, discrete, or continuous, and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Note that y_i is not necessarily assumed to follow an exponential family distribution. Correspondingly, the likelihood function can be written as $L(\beta) = \prod_{i=1}^n f(y_i|x_i^{\mathrm{T}}\beta)$, where $f(\cdot)$ is an unknown probability density function and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of unknown parameters. Also, denote $\mu_i = \int y f(y|x_i^{\mathrm{T}}\beta)dy = E(y_i|x_i^{\mathrm{T}}\beta)$.

Suppose that we have K candidate models with the likelihood function of the kth candidate model being $L_{(k)}(\beta_{(k)}) = \prod_{i=1}^n f_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\beta_{(k)})$, where the function $f_{(k)}$ is known, but it could be misspecified, and $\beta_{(k)}$ is a vector of the unknown parameters. Also, $x_{(k),i}$ is the predictor vector corresponding to the kth model. The maximum likelihood estimator of $\beta_{(k)}$ can be derived by $\hat{\beta}_{(k)} = \arg\max_{\beta_{(k)}} \prod_{i=1}^n f_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\beta_{(k)})$. Then the estimator of μ_i with the observation x_i under the kth model is

$$\hat{\mu}_{(k),i} = \int y f_{(k)}(y|x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)})dy = E_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}),$$

where $E_{(k)}$ is the expectation taken under the kth candidate model. Further, we formulate the model averaging estimator of μ_i as $\hat{\mu}_i(w) = \sum_{k=1}^K w_k\hat{\mu}_{(k),i}$, where $w = (w_1, \ldots, w_K)^{\mathrm{T}}$ and $w \in \mathcal{W} = \{w \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$.

Next, we give a LOOCV criterion to select the model weights. Specifically, let $\hat{\beta}_{(k)}^{-i}$ be the maximum likelihood estimator of $\beta_{(k)}$ in the kth model without using the ith observation. Then the LOOCV estimator of μ_i for the kth model is $\hat{\mu}_{(k)}^{[i]} = \int y f_{(k)}(y|x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}^{-i})dy = E_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}^{-i})$. Denote $\hat{\mu}_{[k]} = (\hat{\mu}_{(k)}^{[1]}, \ldots, \hat{\mu}_{(k)}^{[n]})^{\mathrm{T}}$. The LOOCV weight choice criterion is defined as

$$CV(w) = \|Y - \hat{\tilde{\mu}}(w)\|_2^2, \tag{S.20}$$

where $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\hat{\tilde{\mu}}(w) = \sum_{k=1}^K w_k\hat{\mu}_{[k]}$. Further, the resultant weight estimator is calculated as $\hat{w} = \arg\min_{w\in\mathcal{W}} CV(w)$ with $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_K)^{\mathrm{T}}$.

However, when the sample size is large, it may be a computationally demanding task for the calculation of the weight estimator. If we use GLMs as candidate models to estimate μ_i, then we can employ the approximate LOOCV estimator $\tilde{\beta}_{(k)}^{-i}$ to construct the weight selection criterion. Specifically, the estimator of μ_i for the kth model based on $\tilde{\beta}_{(k)}^{-i}$ is

$$\tilde{\mu}_{(k),i} = \int y f_{(k)}(y|x_{(k),i}^{\mathrm{T}}\tilde{\beta}_{(k)}^{-i})dy = E_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\tilde{\beta}_{(k)}^{-i}).$$

Let $\tilde{\mu}_{[k]} = (\tilde{\mu}_{(k),1}, \ldots, \tilde{\mu}_{(k),n})^{\mathrm{T}}$. The approximate LOOCV weight choice criterion is defined as

$$ACV(w) = \|Y - \tilde{\mu}(w)\|_2^2, \tag{S.21}$$

where $\tilde{\mu}(w) = \sum_{k=1}^K w_k\tilde{\mu}_{[k]}$. Further, the resultant weight estimator is calculated as $\tilde{w} = \arg\min_{w\in\mathcal{W}} ACV(w)$ with $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_K)^{\mathrm{T}}$.

Let $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}} = (E(y_1|x_1^{\mathrm{T}}\beta), \ldots, E(y_n|x_n^{\mathrm{T}}\beta))^{\mathrm{T}}$ and $\hat{\mu}(w) = (\hat{\mu}_1(w), \ldots, \hat{\mu}_n(w))^{\mathrm{T}}$. Next, we present the asymptotic optimality of the proposed averaging estimator $\hat{\mu}(\tilde{w})$ based on the approximate LOOCV weight estimator. Define the risk function as $R(w) = E\|\hat{\mu}(w) - \mu\|_2^2$. The following conditions are required for the asymptotic optimality.

Condition S3 *Suppose that $K \le n$. There exists a limiting value $\beta_{(k)}^*$ for $\hat{\beta}_{(k)}$ such that $\hat{\beta}_{(k)} - \beta_{(k)}^* = O_p(K^{1/2}n^{-1/2})$ uniformly for $k = 1, \ldots, K$.*

Condition S4 *For $i = 1, \ldots, n$, (i) $E(\varepsilon_i^2)$ and $E(y_i^2)$ are both $O(1)$, where $\varepsilon_i = y_i - \mu_i$; (ii) $\tilde{\mu}_{(k),i}$ is differentiable with respect to $\tilde{\beta}_{(k)}^{-i}$; (iii) there exists a constant $\delta^* > 0$ such that $E\big(\sup_{\beta_{(k)}\in\mathcal{B}(\beta_{(k)}^*|\,\delta^*)} \big\|\frac{\partial\tilde{\mu}_{(k),i}}{\partial\tilde{\beta}_{(k)}^{-i}}\big|_{\tilde{\beta}_{(k)}^{-i}=\beta_{(k)}}\big\|_2^2\big) = O(1)$, and $\mu_{(k),i}^*$ is $O(1)$ uniformly for $k = 1, \ldots, K$ and $i = 1, \ldots, n$, where $\mu_{(k),i}^* = E_{(k)}(y_i|x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*)$.*

Condition S5 $\xi_n^{-1}\sup_{w\in\mathcal{W}}[|\|\hat{\mu}(w)-\mu\|_2^2-\|\mu^*(w)-\mu\|_2^2|]$ *is uniformly integrable, where* $\xi_n=\inf_{w\in\mathcal{W}}R^*(w)$ *with* $R^*(w)=\|\mu^*(w)-\mu\|_2^2$ *and* $\mu^*(w)=\sum_{k=1}^K w_k\mu^*_{(k),i}$.

Condition S6 $\mathrm{Var}(\mu^*_{(k),i}\varepsilon_i)$ *are bounded by a constant uniformly for* $k=1,\ldots,K$.

Condition S7 $n^{1/2}K\xi_n^{-1}=o(1)$.

Conditions S3–S4 and S6 are quite similar to Assumptions 1–2 and 4 in Zhang and Liu (2023). Thus, for a justification of the rationality of these conditions, see Zhang and Liu (2023). Condition S5 is only imposed to ensure that the expectation of the above equation is $o(1)$. Condition S7 puts a bound on the number of models relative to the sample size. This condition is similar to Condition C.6 of Zhang et al. (2016) and Condition A3 of Ando and Li (2017). Note that Condition S3 implies that the dimension of $\beta_{(k)}$ is fixed. Hence, the following asymptotic optimality result is derived under the assumption that the dimensions of the candidate models are fixed.

THEOREM S.2. *Under Conditions S3–S7 and the conditions of Lemma 1, we have* $R(\tilde{w})/\inf_{w\in\mathcal{W}}R(w)\xrightarrow{P}1$.

Theorem S.2 shows that the proposed averaging estimator $\hat{\mu}(\tilde{w})$ remains asymptotically optimal in the sense that minimizing the expected squared error $R(w)$ even if the true density function of y_i does not belong to the exponential family.

Proof of Theorem S.2. Let $ACV^*(w)=ACV(w)-(Y-\mu)^{\mathrm{T}}(Y+\mu)$, where $Y=(y_1,\ldots,y_n)^{\mathrm{T}}$. Thus, $\tilde{w}=\arg\min_{w\in\mathcal{W}}ACV(w)=\arg\min_{w\in\mathcal{W}}ACV^*(w)$. According to Zhang and Liu (2023), Theorem S.2 is valid if the following hold:

$$\sup_{w\in\mathcal{W}}\frac{|R(w)-R^*(w)|}{R^*(w)}=o(1)\text{ and }\sup_{w\in\mathcal{W}}\frac{|ACV^*(w)-R^*(w)|}{R^*(w)}=o_p(1). \tag{S.22}$$

We first focus on the first term of (S.22). Note that by Conditions S3–S4 and S7,

$$\begin{aligned}
&\xi_n^{-1}\sup_{w\in\mathcal{W}}\sum_{i=1}^n|(\hat{\mu}_i(w)-\mu_i)^2-(\mu_i^*(w)-\mu_i)^2|=\xi_n^{-1}\sup_{w\in\mathcal{W}}\sum_{i=1}^n|(\hat{\mu}_i(w)-\mu_i^*(w))(\hat{\mu}_i(w)-\mu_i^*(w)+2\mu_i^*(w)-2\mu_i)|\\
&=\xi_n^{-1}\sup_{w\in\mathcal{W}}\sum_{i=1}^n\left|\left(\sum_{k=1}^K w_k(\hat{\beta}_{(k)}-\beta^*_{(k)})^{\mathrm{T}}\frac{\partial\hat{\mu}_{(k),i}}{\partial\hat{\beta}_{(k)}}\Big|_{\hat{\beta}_{(k)}=\bar{\beta}_{(k)}}\right)\sum_{k=1}^K w_k\left((\hat{\beta}_{(k)}-\beta^*_{(k)})^{\mathrm{T}}\frac{\partial\hat{\mu}_{(k),i}}{\partial\hat{\beta}_{(k)}}\Big|_{\hat{\beta}_{(k)}=\bar{\beta}_{(k)}}+2\mu^*_{(k),i}-2\mu_i\right)\right|\\
&=\xi_n^{-1}O_p(n^{1/2}K^{1/2})=o_p(1),
\end{aligned} \tag{S.23}$$

where $\bar{\beta}_{(k)}$ lies between $\hat{\beta}_{(k)}$ and $\beta^*_{(k)}$. Hence, with (S.23), we have

$$\sup_{w\in\mathcal{W}}\frac{|R(w)-R^*(w)|}{R^*(w)}\le E\left(\xi_n^{-1}\sup_{w\in\mathcal{W}}|\|\hat{\mu}(w)-\mu\|_2^2-\|\mu^*(w)-\mu\|_2^2|\right)=o(1), \tag{S.24}$$

where the third step uses Condition S5. This finishes the proof of the first term of (S.22).

We next consider the second term of (S.22). Under the conditions of Lemma 1, we have $\max_{1\le i\le n}\max_{1\le k\le K}\|\tilde{\beta}^{-i}_{(k)}-\hat{\beta}_{(k)}\|_2=O_p(n^{-1}\log(n))$, which implies that $\max_{1\le i\le n}\max_{1\le k\le K}\|\tilde{\beta}^{-i}_{(k)}-\beta^*_{(k)}\|_2=O_p(n^{-1/2}K^{1/2})$ based on Condition S3. Note that

$$\begin{aligned}
&|ACV^*(w)-R^*(w)|=|[\|Y-\tilde{\mu}(w)\|_2^2-(Y-\mu)^{\mathrm{T}}(Y+\mu)]-\|\mu^*(w)-\mu\|_2^2|\\
&\le|[\|Y-\mu^*(w)\|_2^2-(Y-\mu)^{\mathrm{T}}(Y+\mu)]-\|\mu^*(w)-\mu\|_2^2||\|Y-\tilde{\mu}(w)\|_2^2-\|Y-\mu^*(w)\|_2^2|/n\\
&=|[\|\mu^*(w)-\mu+\mu-Y\|_2^2-(Y-\mu)^{\mathrm{T}}(Y+\mu)]-\|\mu^*(w)-\mu\|_2^2||\|Y-\tilde{\mu}(w)\|_2^2-\|Y-\mu^*(w)\|_2^2|\\
&\le 2|\mu^*(w)^{\mathrm{T}}(\mu-Y)|+|\|Y-\tilde{\mu}(w)\|_2^2-\|Y-\mu^*(w)\|_2^2|.
\end{aligned}$$

Denote $\tilde{\mu}_i(w) = \sum_{k=1}^K w_k \tilde{\mu}_{(k),i}$ and $\mu_i^*(w) = \sum_{k=1}^K w_k \mu_{(k),i}^*$. Similar to (S.23), by Conditions S3 and S4, we have

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} |\|Y-\tilde{\mu}(w)\|_2^2 - \|Y-\mu^*(w)\|_2^2| = \sup_{w\in\mathcal{W}} \left| \sum_{i=1}^n \left[(\tilde{\mu}_i(w)-y_i)^2 - (\mu_i^*(w)-y_i)^2 \right] \right| \\
&= \sup_{w\in\mathcal{W}} \left| \sum_{i=1}^n \left[(\tilde{\mu}_i(w)-\mu_i^*(w))(\tilde{\mu}_i(w)+\mu_i^*(w)-2y_i) \right] \right| \\
&\le \sup_{w\in\mathcal{W}} \left| \sum_{i=1}^n \left(\sum_{k=1}^K w_k \|\tilde{\beta}_{(k)}^{-i} - \beta_{(k)}^*\|_2 \left\| \frac{\partial \tilde{\mu}_{(k),i}}{\partial \tilde{\beta}_{(k)}^{-i}} \Big|_{\tilde{\beta}_{(k)}^{-i}=\bar{\beta}_{(k)i}} \right\|_2 \right) \left(\sum_{k=1}^K w_k \|\tilde{\beta}_{(k)}^{-i} - \beta_{(k)}^*)\|_2 \right. \right. \\
&\quad \left. \left. \left\| \frac{\partial \tilde{\mu}_{(k),i}}{\partial \tilde{\beta}_{(k)}^{-i}} \Big|_{\tilde{\beta}_{(k)}^{-i}=\bar{\beta}_{(k)i}} \right\|_2 + 2\sum_{k=1}^K w_k |\mu_{(k),i}^*| + 2|\mu_i| + 2|\varepsilon_i| \right) \right| = O_p(K) + O_p(n^{1/2}K^{1/2}), \qquad \text{(S.25)}
\end{aligned}$$

where $\bar{\beta}_{(k)i}$ is in $\mathcal{B}(\beta_{(k)}^*|\delta^*)$ for all i.

Then for any $\epsilon > 0$, using the Chebyshev's inequality,

$$P\left(\xi_n^{-1} \sup_{w\in\mathcal{W}} |\mu^*(w)^{\mathrm{T}}(\mu - Y)| > \epsilon \right) \le \sum_{k=1}^K P\left(\left| \sum_{i=1}^n \mu_{(k),i}^*(y_i-\mu_i) \right| > \epsilon\xi_n \right) \le \xi_n^{-2}\epsilon^{-2} n \sum_{k=1}^K \mathrm{Var}(\mu_{(k),i}^*\varepsilon_i). \qquad \text{(S.26)}$$

By (S.25)–(S.26) and Conditions S6–S7, we obtain the second term of (S.22). This completes the proof.

In the following, we present a simulation study to evaluate the performance of our method. Specifically, we generate y_i by a negative binomial distribution

$$\Pr(y_i = k) = \binom{k+r_i-1}{k} p^k (1-p)^{r_i}, \text{ for } k = 0, 1, 2, \ldots,$$

where $p = 0.5$ is the success probability in each experiment, $r_i = [2\exp(x_i^{\mathrm{T}}\beta)]$ is the number of failures until the experiment is stopped, and $\beta = (0.3, 0.15, -0.6, 0, -0.2, -0.05, 0.5, 0, 0, \ldots, 0)_p^{\mathrm{T}}$ with $p = [3n^{1/3}]$. Here $[x]$ is the smallest integer larger than x. We adopt the DCMS procedure to prepare candidate models. Under each candidate model, we take Poisson regression to fit the data. Thus the candidate models are still misspecified, because the response follows the negative binomial distribution instead of the Poisson distribution. To assess the performance of each method, we independently generate a testing data $\{x_s, y_i\}_{s=1}^{100}$ for the rth replication and calculate $\mathrm{MSE}^{(r)} = 100^{-1}\sum_{s=1}^{100}(\hat{E}(y_s|x_s) - E(y_s|x_s))^2$, where $\hat{E}(y_s|x_s)$ is obtained by model selection or averaging methods. Further, we use $\mathrm{MSE} = R^{-1}\sum_{r=1}^R \mathrm{MSE}^{(r)}$ to evaluate each method based on R replications. We set the sample size $n \in \{100, 200, 300, 400, 800\}$ and $R = 200$.

Simulation results are shown in Table S9. JMA and EJMA calculate the weight estimators by (S.20) and (S.21), respectively. Similar to JMA, the weight estimators for $\mathrm{FJMA_S}$ and MACV are also obtained under the squared loss. It can be seen from Table S9 that EJMA significantly outperforms AIC, BIC, SAIC, SBIC and EMA in terms of MSE. Moreover, $\mathrm{OMA_1}$ and $\mathrm{OMA_2}$ always yield larger MSE values than EJMA. Although JMA, EJMA, $\mathrm{FJMA_S}$ and MACV exhibit comparable performance concerning MSE, EJMA stands out as the most computationally efficient. In addition, as the sample size increases from 100 to 800, the MSE_w value of EJMA decreases from 0.0001 to 5e-7, which demonstrates the high approximation accuracy of the weight estimator of EJMA.

Part G: Extension to High-Dimensional GLMs

In this section, we extend our method to the high-dimensional case, that is, the dimension p is larger than the sample size n. But we still require that the number of covariates with nonzero coefficients in GLMs does not exceed the sample size, i.e., $p^* < n$ with $p^* = |\mathcal{I}_T|$. Since $p > n$, it is impossible and computationally infeasible to consider all-subset models as candidate models. In this situation, we average some models whose dimensions are smaller than n. This idea is also frequently used in the high-dimensional model averaging studies (see, e.g., Ando and Li (2017) and Zhang et al. (2020)).

Table S9 The simulation results in terms of MSE and the computational time of various methods for the negative binomial distribution.

MSE		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
$n=100$	Mean	18.956	16.769	18.185	15.916	12.482	13.367	11.700	11.124	11.122	11.119	11.155
	SE	(1.267)	(1.177)	(1.272)	(1.152)	(0.405)	(0.862)	(0.517)	(0.405)	(0.405)	(0.407)	(0.401)
$n=200$	Mean	10.128	11.517	10.038	11.176	10.024	7.890	8.154	7.639	7.638	7.632	7.702
	SE	(0.363)	(0.381)	(0.363)	(0.387)	(0.295)	(0.275)	(0.277)	(0.258)	(0.258)	(0.258)	(0.258)
$n=300$	Mean	7.326	9.146	7.327	8.715	8.496	5.854	6.089	5.607	5.607	5.604	5.666
	SE	(0.239)	(0.243)	(0.239)	(0.250)	(0.211)	(0.190)	(0.213)	(0.187)	(0.187)	(0.187)	(0.192)
$n=400$	Mean	5.985	8.314	5.982	8.002	8.236	4.942	5.313	4.740	4.740	4.738	4.794
	SE	(0.180)	(0.210)	(0.180)	(0.204)	(0.177)	(0.146)	(0.161)	(0.133)	(0.133)	(0.133)	(0.137)
$n=800$	Mean	3.954	5.287	3.954	5.126	7.542	3.554	3.877	3.495	3.495	3.495	3.527
	SE	(0.092)	(0.195)	(0.092)	(0.180)	(0.168)	(0.083)	(0.010)	(0.083)	(0.083)	(0.083)	(0.083)
Time		AIC	BIC	SAIC	SBIC	EMA	OMA_1	OMA_2	$FJMA_S$	JMA	EJMA	MACV
$n=100$	Mean	0.115	0.115	0.115	0.115	0.115	0.300	0.304	0.974	3.618	0.276	0.577
$n=200$	Mean	0.300	0.300	0.300	0.300	0.300	0.623	0.628	2.818	11.866	0.632	1.107
$n=300$	Mean	0.553	0.553	0.553	0.553	0.553	0.999	1.013	5.463	24.148	1.047	1.721
$n=400$	Mean	0.846	0.846	0.846	0.846	0.846	1.331	1.357	7.837	39.161	1.470	2.256
$n=800$	Mean	2.505	2.505	2.505	2.505	2.505	3.321	3.302	21.083	135.708	3.760	5.107

Note: All MSE values has been magnified by a factor of 10. The computing time are obtained in R×64 4.1.2 by a computer with an Intel(R) Core(TM) i7-8550U CPU (1.80GHz) with 512 GB memory.

Specifically, we can rank the covariates in descending order based on their distance correlation coefficients with the response variable. Then we construct K nested models as the candidate model set $\mathcal{M}_H$, where the kth model contains the first k covariates according to the order. Let $K < n$ with $K = \bar{p} = \max_{1\le k\le K} p_k$. The idea of using nested candidate models is also advocated by Zhang et al. (2016) and Zhang et al. (2020). Under the kth candidate model, we calculate the approximate jackknife estimator $\tilde{\theta}_{(k),i}$. Further, based on the candidate model set $\mathcal{M}_H$, we select the weight vector by $\tilde{w}_H = \arg\max_{w\in\mathcal{W}} ACV(w)$ with $\tilde{w}_H = (\tilde{w}_{H,1},\ldots,\tilde{w}_{H,K})^{\mathrm{T}}$. The corresponding model averaging estimator of θ_i is $\hat{\theta}_i(\tilde{w}_H) = \sum_{k=1}^{K} \tilde{w}_{H,k} x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)}$.

Next, we study the property of the model averaging estimator in the high-dimensional situation.

THEOREM S.3. *Under Conditions 1–6, we have* $KL(\tilde{w}_H)/\inf_{w\in\mathcal{W}} KL(w) \xrightarrow{P} 1$ *as* $n\to\infty$. *Further, under Conditions 1–7, we have* $R(\tilde{w}_H)/\inf_{w\in\mathcal{W}} R(w) \xrightarrow{P} 1$ *as* $n\to\infty$, *where* $R(w) = E_y(KL(w))$.

Theorem S.3 shows that the asymptotic optimality of EJMA estimator still holds for the high-dimensional situation. Note that the conditions of Theorem S.3 are the same as those of Theorem 1, which is because these conditions are all related to the given candidate models and p is not used in the proof of Theorem S.3.

In the following, we provide some simulations to evaluate the performance of the proposed model averaging method. We consider the following logistic regression model:

$$\operatorname{logit}\{\Pr(y_i=1)\} = \sum_{j=1}^{p} x_{ij}\beta_j = x_i^{\mathrm{T}}\beta, \quad i=1,\ldots,n, \tag{S.27}$$

where $x_i = (x_{i1},\ldots,x_{ip})$ are generated from the multivariate normal distribution $N_p(0,\Sigma)$ with $\Sigma = (\rho^{|i-j|})_{p\times p}$ and $\rho = 0.8$. The number of nonzero elements β_j with $\beta_j \neq 0$ is $p^* = [2n^{1/3}]$, where $[x]$ denotes the smallest integer greater than x. Following Ando and Li (2017), we generate these nonzero β_j from standard normal distribution $N(0,1)$. Let the nonzero β_j be evenly spaced, where $j = 10(j^*-1)+1$ and $j^* = 1,\ldots,p^*$. To mimic the case of model misspecification, the variable x_{i1} is missed when producing the candidate models. The number of candidate models is $K = [3n^{1/3}]$. Four cases of (n,p) are considered: $(100,200)$, $(200,400)$, $(300,600)$ and $(400,800)$. In addition, we set $n_{\text{test}} = 100$ and the number of replications $R = 200$.

In this simulation, we also compare our method EJMA with the commonly used regularization techniques, including the least absolute shrinkage and selection operator (LASSO; Tibshirani (1996)), the smoothly clipped absolute deviation (SCAD; Fan and Li (2001)) and the minimax concave penalty (MCP; Zhang

(2010)). The optimal tunning parameters in LASSO, SCAD and MCP are chosen by 5-fold cross-validation, which can be implemented by R package `ncvreg`.

Table S10 reports the KL loss of various methods for the high-dimensional logistic model (S.27). It is seen that EJMA yields smaller KL loss values than AIC, BIC, SAIC, SBIC, EMA, LASSO, SCAD and MCP in all cases. And EJMA, JMA and MACV perform almost equally well. Moreover, we calculate the MSE_w values of weight estimators in four cases of (n, p), which are 0.00303, 0.00021, 0.00007 and 0.00005, respectively. It can be observed that the MSE_w value decreases with the increase of sample size, which illustrates that the EJMA weight estimator is still a satisfactory approximation to the JMA weight estimator in the high-dimensional situation especially with the large sample size. Also, we compare the computational time of JMA, MACV and EJMA in Table S11 for the high-dimensional logistic model (S.27). As shown in Table S11, JMA has a serious computational burden especially when the sample size is large. In contrast, EJMA has a satisfactory computational efficiency which takes much less computing time relative to JMA. Also, EJMA is superior to MACV in terms of computational speed.

Table S10 The simulation results of various methods in terms of KL loss for the high-dimensional logistic model.

High Dimension		AIC	BIC	SAIC	SBIC	EMA	LASSO	SCAD
$(n, p) = (100, 200)$	Mean	4.916	2.872	4.413	2.730	2.871	9.231	11.678
	SE	(0.199)	(0.082)	(0.178)	(0.074)	(0.074)	(0.857)	(1.414)
$(n, p) = (200, 400)$	Mean	4.981	5.061	4.905	4.900	4.405	30.927	12.688
	SE	(0.091)	(0.087)	(0.089)	(0.082)	(0.065)	(4.365)	(0.884)
$(n, p) = (300, 600)$	Mean	5.037	4.961	4.960	4.866	4.681	19.190	10.213
	SE	(0.076)	(0.072)	(0.074)	(0.070)	(0.060)	(1.866)	(0.580)
$(n, p) = (400, 800)$	Mean	2.678	2.760	2.637	2.702	2.679	28.507	8.492
	SE	(0.050)	(0.050)	(0.048)	(0.049)	(0.041)	(4.168)	(0.817)
High Dimension		MCP	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
$(n, p) = (100, 200)$	Mean	5.685	3.197	2.578	2.749	2.586	2.597	2.601
	SE	(0.371)	(0.100)	(0.059)	(0.077)	(0.067)	(0.072)	(0.064)
$(n, p) = (200, 400)$	Mean	6.117	4.372	4.285	4.300	4.276	4.279	4.271
	SE	(0.283)	(0.071)	(0.066)	(0.069)	(0.068)	(0.069)	(0.067)
$(n, p) = (300, 600)$	Mean	5.403	4.605	4.543	4.548	4.537	4.539	4.535
	SE	(0.141)	(0.063)	(0.057)	(0.062)	(0.062)	(0.062)	(0.062)
$(n, p) = (400, 800)$	Mean	3.159	2.488	2.576	2.461	2.459	2.460	2.462
	SE	(0.102)	(0.043)	(0.039)	(0.042)	(0.042)	(0.042)	(0.042)

Note: All values has been magnified by a factor of 10.

Table S11 The average computational time of various methods for the high-dimensional logistic model.

High Dimension	AIC	BIC	SAIC	SBIC	EMA	LASSO	SCAD
$(n, p) = (100, 200)$	0.115	0.115	0.115	0.115	0.115	0.415	1.549
$(n, p) = (200, 400)$	0.572	0.572	0.572	0.572	0.572	1.315	4.527
$(n, p) = (300, 600)$	1.639	1.639	1.639	1.639	1.639	2.884	9.252
$(n, p) = (400, 800)$	3.326	3.326	3.326	3.326	3.326	4.716	10.034
High Dimension	MCP	OMA_1	OMA_2	FJMA_S	JMA	EJMA	MACV
$(n, p) = (100, 200)$	1.250	0.353	0.361	2.504	4.575	0.350	0.665
$(n, p) = (200, 400)$	3.513	0.957	0.969	7.156	15.821	0.943	1.541
$(n, p) = (300, 600)$	7.140	2.140	2.152	13.731	34.164	2.131	2.986
$(n, p) = (400, 800)$	10.503	3.906	3.912	21.621	58.011	3.903	4.931

Note: The simulations are conducted in R×64 4.1.2 by a computer with an Intel(R) Core(TM) i7-8550U CPU (1.80GHz) with 512 GB memory.

Part H: Proofs of the Theoretical Results

To prove the asymptotic optimality theory, the following lemma is developed.

LEMMA S.1. *If Condition 6 and*

$$\sup_{w\in\mathcal{W}} |AL(w)/R(w) - 1| \xrightarrow{P} 0, \ \sup_{w\in\mathcal{W}} |AP(w)/KL^*(w)| \xrightarrow{P} 0, \ \textit{and} \ \sup_{w\in\mathcal{W}} |R(w)/KL^*(w) - 1| \to 0 \quad \text{(S.28)}$$

hold, where $AL(w) = \sum_{i=1}^{n} \phi^{-1}(b'(\theta_i)(\theta_i - \sum_{k=1}^{K} w_k \tilde{\theta}_{(k),i}) - (b(\theta_i) - b(\sum_{k=1}^{K} w_k \tilde{\theta}_{(k),i})))$ *and* $AP(w) = \sum_{i=1}^{n}(\phi^{-1}\varepsilon_i(\sum_{k=1}^{K} w_k \tilde{\theta}_{(k),i} - \theta_i)$*, then* $R(\tilde{w})/\inf_{w\in\mathcal{W}} R(w) \xrightarrow{P} 1$ *as* $n \to \infty$*, where* $R(w) = E_y(KL(w))$ *and* $\tilde{w} = \arg\max_{w\in\mathcal{W}} ACV(w)$.

H.1. Proof of Lemma S.1

Note that

$$-ACV(w) = AL(w) - AP(w) - \sum_{i=1}^{n}(\phi^{-1}\varepsilon_i\theta_i + \phi^{-1}b'(\theta_i)\theta_i - \phi^{-1}b(\theta_i) + c(y_i,\phi)). \tag{S.29}$$

It is observed that the third term in (S.29) is unrelated to w. Hence, $\tilde{w} = \arg\min_{w\in\mathcal{W}} -ACV(w) = \arg\min_{w\in\mathcal{W}}(AL(w) - AP(w))$. Based on the definition of infimum, there exist a weight vector sequences $\{w^n\}$ ($w^n = (w_1^n, \ldots, w_K^n) \in \mathcal{W}$) and a sequence $\{\varpi_n\}$ such that $\varpi_n \to 0$ and $\inf_{w\in\mathcal{W}} R(w) = R(w^n) - \varpi_n$ as $n \to \infty$. It is readily shown that $R(w^n)/\inf_{w\in\mathcal{W}} R(w) > 1$, and by Condition 6 and the third term of (S.28),

$$\frac{\varpi_n}{\inf_{w\in\mathcal{W}} R(w)} = \frac{\varpi_n/KL^*(w^n)}{(R(w^n) - \varpi_n)/KL^*(w^n)} = \frac{\varpi_n/KL^*(w^n)}{R(w^n)/KL^*(w^n) - \varpi_n/KL^*(w^n)} \to 0.$$

Denote $ALP(w) = AL(w) - AP(w)$. It is observed that for any $\epsilon > 0$,

$$P\left(\left|\frac{\inf_{w\in\mathcal{W}} R(w)}{R(\tilde{w})} - 1\right| > \epsilon\right) \le P\left(\frac{R(\tilde{w}) - ALP(\tilde{w}) + ALP(w^n) - R(w^n) + \varpi_n}{R(\tilde{w})} > \epsilon\right) \triangleq P_1.$$

Notice that by (S.28), $\frac{R(\tilde{w})-ALP(\tilde{w})}{R(\tilde{w})} \le \sup_{w\in\mathcal{W}}\left|\frac{R(w)-AL(w)}{R(w)}\right| + \frac{\sup_{w\in\mathcal{W}}|AP(w)/KL^*(w)|}{(R(\tilde{w})/KL^*(\tilde{w})-1)+1} \xrightarrow{P} 0$. Moreover, according to Condition 6 and (S.28),

$$\begin{aligned}\frac{ALP(w^n) - R(w^n) + \varpi_n}{R(\tilde{w})} &\le \frac{|AL(w^n) - R(w^n)|}{\inf_{w\in\mathcal{W}} R(w)} + \frac{|AP(w^n)/KL^*(w^n)|}{\inf_{w\in\mathcal{W}} R(w)/KL^*(w^n)} + \frac{\varpi_n}{\inf_{w\in\mathcal{W}} R(w)} \\ &\le \frac{|AL(w^n) - R(w^n)|/R(w^n)}{1 - \varpi_n/R(w^n)} + \frac{|AP(w^n)/KL^*(w^n)|}{(R(w^n) - \varpi_n)/KL^*(w^n)} + \frac{\varpi_n}{\inf_{w\in\mathcal{W}} R(w)} \xrightarrow{P} 0.\end{aligned}$$

Thus, we have $P_1 \to 0$, which further implies that $P(|\inf_{w\in\mathcal{W}} R(w)/R(\tilde{w}) - 1| > \epsilon) \to 0$. This completes the proof of Lemma S.1. □

H.2. Proof of Lemma 1

First, we give the proof of (14). The log-likelihood function under the kth candidate model can be written as

$$F_{(k)}(\beta_{(k)}) = \sum_{i=1}^{n} \phi^{-1}(y_i\theta_{(k),i} - b(\theta_{(k),i})) + c(y_i,\phi), \tag{S.30}$$

where $\theta_{(k),i} = x_{(k),i}^{\mathrm{T}}\beta_{(k)}$. Further, it is readily shown that $\frac{\partial F_{(k)}(\beta_{(k)})}{\partial \beta_{(k)}}\big|_{\beta_{(k)}=\hat{\beta}_{(k)}} = \sum_{i=1}^{n} \phi^{-1}(y_i - b(x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}))x_{(k),i} = 0$.

Let $\mathcal{N}(\beta_{(k)}^*|\delta) = \{\beta_{(k)} \in \mathcal{R}^{p_k} : (K\bar{p})^{-1/2}n^{1/2}\|\beta_{(k)} - \beta_{(k)}^*\|_2 \le \delta\}$ and $\bar{\mathcal{N}}(\beta_{(k)}^*|\delta)$ be the boundary of $\mathcal{N}(\beta_{(k)}^*|\delta)$ for some fixed $\delta > 0$. Note that $(K\bar{p})^{1/2}n^{-1/2}\delta < \delta^*$ for a sufficiently large n, where δ^* is defined in Conditions 1 and 3. Denote $\theta_{(k),i}^* = x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*$. Using the Taylor expansion of (S.30), there exists a large $\delta > 0$ such that when n is sufficiently large, $\beta_{(k)} \in \bar{\mathcal{N}}(\beta_{(k)}^*|\delta)$, and

$$\begin{aligned}\max_{1\le k\le K}(F_{(k)}(\beta_{(k)}) - F_{(k)}(\beta_{(k)}^*)) = \max_{1\le k\le K}\sum_{i=1}^{n} \phi^{-1}\big((y_i x_{(k),i}^{\mathrm{T}}(\beta_{(k)} - \beta_{(k)}^*) - b'(\theta_{(k),i}^*)x_{(k),i}^{\mathrm{T}}(\beta_{(k)} - \beta_{(k)}^*) \\ - \frac{1}{2}b''(\bar{\theta}_{(k),i}^*)(\beta_{(k)} - \beta_{(k)}^*)^{\mathrm{T}}x_{(k),i}x_{(k),i}^{\mathrm{T}}(\beta_{(k)} - \beta_{(k)}^*)\big) \triangleq T_1,\end{aligned}$$

where $\bar{\theta}^*_{(k),i} = x^{\mathrm{T}}_{(k),i}\bar{\beta}^*_{(k)}$ and $\bar{\beta}^*_{(k)}$ lies between $\beta_{(k)}$ and $\beta^*_{(k)}$. By the Schwarz inequality and Condition 3, T_1 can be calculated as

$$\begin{aligned} T_1 \leq & \max_{1\leq k\leq K} \phi^{-1}\left\| \sum_{i=1}^{n}(y_i - b'(\theta^*_{(k),i}))x_{(k),i}\right\|_2 \|\beta_{(k)} - \beta^*_{(k)}\|_2 - \frac{n}{2\phi}\min_{1\leq k\leq K}\|\beta_{(k)} - \beta^*_{(k)}\|_2^2 \min_{1\leq k\leq K}\lambda_{\min}(\mathcal{H}_k(\bar{\beta}^*_{(k)})) \\ \leq & \phi^{-1}\frac{\delta\sqrt{K\bar{p}}}{\sqrt{n}}\max_{1\leq k\leq K}\left\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\right\|_2 - C\delta^2 K\bar{p}, \end{aligned}$$

where $\varepsilon_{(k),i} = y_i - b'(\theta^*_{(k),i})$. We next show that $\max_{1\leq k\leq K}\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\|_2 = O_p(\sqrt{nK\bar{p}})$. By the Markov's inequality, we have

$$P\left(\max_{1\leq k\leq K}\frac{1}{\sqrt{nK\bar{p}}}\left\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\right\|_2 > M\right) \leq \sum_{k=1}^{K} P\left(\left\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\right\|_2 > M\sqrt{nK\bar{p}}\right) \leq \sum_{k=1}^{K}\frac{E\|\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}\|_2^2}{K\bar{p}nM^2},$$

where $\varepsilon_{(k)} = (\varepsilon_{(k),1},\ldots,\varepsilon_{(k),n})^{\mathrm{T}}$. By the elementary calculations, we derive that $E\|\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}\|_2^2 = E(\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}X^{\mathrm{T}}_{(k)}\varepsilon_{(k)}) = tr(X_{(k)}X^{\mathrm{T}}_{(k)}E(\varepsilon_{(k)}\varepsilon^{\mathrm{T}}_{(k)}))$ and further we have

$$E\|\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}\|_2^2 = \sum_{i=1}^{n}\|x_{(k),i}\|_2^2\mathrm{Var}(\varepsilon_{(k),i}) + \sum_{i=1}^{n}\sum_{j=1}^{n}x^{\mathrm{T}}_{(k),i}x_{(k),j}E(\varepsilon_{(k),i})E(\varepsilon_{(k),j}), \tag{S.31}$$

where $tr(\cdot)$ is the trace of a matrix. Note that the KL divergence between the kth candidate model and the true model is $KL(\beta_{(k)}) = \sum_{i=1}^{n}\phi^{-1}(b'(\theta_i)(\theta_i - \theta_{(k),i}) - (b(\theta_i) - b(\theta_{(k),i})))$. Further, by the definition of $\beta^*_{(k)}$, we have $\frac{\partial KL(\beta_{(k)})}{\partial\beta_{(k)}}\big|_{\beta_{(k)}=\beta^*_{(k)}} = \sum_{i=1}^{n}\phi^{-1}(b'(\theta^*_{(k),i}) - b'(\theta_i))x_{(k),i} = 0$. Thus,

$$E\left(\sum_{j=1}^{n}\varepsilon_{(k),j}x_{(k),j}\right) = E\left(\sum_{j=1}^{n}(y_j - b'(\theta_j))x_{(k),j}\right) + \sum_{i=1}^{n}(b'(\theta_i) - b'(\theta^*_{(k),i}))x_{(k),i} = 0.$$

Then, by Conditions 1 and 2, $E\|\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}\|_2^2$ is calculated as $E\|\varepsilon^{\mathrm{T}}_{(k)}X_{(k)}\|_2^2 = \sum_{i=1}^{n}\|x_{(k),i}\|_2^2\mathrm{Var}(\varepsilon_{(k),i}) \leq \max_{1\leq i\leq n}\max_{1\leq k\leq K}\|x_{(k),i}\|_2^2\sum_{i=1}^{n}\mathrm{Var}(y_i) \leq C\bar{p}n$. Hence, as $M\to\infty$,

$$P\left(\max_{1\leq k\leq K}(nK\bar{p})^{-1/2}\left\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\right\|_2 > M\right) \leq \sum_{k=1}^{K}\frac{C\bar{p}n}{K\bar{p}nM^2} = \frac{CK\bar{p}n}{K\bar{p}nM^2} \to 0.$$

Further, we deduce that $\max_{1\leq k\leq K}\| \sum_{i=1}^{n}\varepsilon_{(k),i}x_{(k),i}\|_2 = O_p(\sqrt{nK\bar{p}})$. In the end, for sufficient large n,

$$\max_{1\leq k\leq K}(F_{(k)}(\beta_{(k)}) - F_{(k)}(\beta^*_{(k)})) \leq K\bar{p}(\phi^{-1}\delta O_p(1) - C\delta^2).$$

For a sufficiently large δ, the first term $\phi^{-1}\delta O_p(1)$ is dominated by the second term $C\delta^2$. Additionally, by Condition 3, for $\beta_{(k)} \in \mathcal{N}(\beta^*_{(k)}|\delta)$, we know that $\frac{\partial^2 F_{(k)}(\beta_{(k)})}{\partial\beta_{(k)}\partial\beta^{\mathrm{T}}_{(k)}} = -\sum_{i=1}^{n}\phi^{-1}b''(\theta_{(k),i})x_{(k),i}x^{\mathrm{T}}_{(k),i} < 0$, that is, the log-likelihood function $F_{(k)}(\beta_{(k)})$ is concave. Therefore, the conclusion (14) is derived.

Second, we provide the proof of (15). Recall that $F^{-i}_{(k)}(\beta_{(k)}) = \sum_{j\neq i}\phi^{-1}(y_j\theta_{(k),j} - b(\theta_{(k),j})) + c(y_j,\phi)$. Denote $\mathcal{N}_1(\hat{\beta}_{(k)}|\delta) = \{\beta_{(k)} \in \mathcal{R}^{p_k} : (\log(n))^{-1}\bar{p}^{-1/2}n\|\beta_{(k)} - \hat{\beta}_{(k)}\|_2 \leq \delta\}$ and let $\bar{\mathcal{N}}_1(\hat{\beta}_{(k)}|\delta)$ be the the boundary of $\mathcal{N}_1(\hat{\beta}_{(k)}|\delta)$. By Taylor expansion of $F^{-i}_{(k)}(\beta_{(k)})$ at $\hat{\beta}_{(k)}$, for $\beta_{(k)} \in \bar{\mathcal{N}}_1(\hat{\beta}_{(k)}|\delta)$ and a sufficiently large n, we have

$$\max_{1\leq i\leq n}\max_{1\leq k\leq K}(F^{-i}_{(k)}(\beta_{(k)}) - F^{-i}_{(k)}(\hat{\beta}_{(k)}))$$

$$= \max_{1\le i\le n}\max_{1\le k\le K}\frac{\partial F_{(k)}^{-i}(\hat\beta_{(k)})}{\partial\beta_{(k)}}(\beta_{(k)}-\hat\beta_{(k)})+\frac{1}{2}(\beta_{(k)}-\hat\beta_{(k)})^{\mathrm{T}}\frac{\partial^2 F_{(k)}^{-i}(\bar\beta_{(k)})}{\partial\beta_{(k)}\partial\beta_{(k)}^{\mathrm{T}}}(\beta_{(k)}-\hat\beta_{(k)})$$

$$= \max_{1\le i\le n}\max_{1\le k\le K}\phi^{-1}(y_i-b'(\hat\theta_{(k),i}))x_{(k),i}^{\mathrm{T}}(\beta_{(k)}-\hat\beta_{(k)})+\frac{1}{2}(\beta_{(k)}-\hat\beta_{(k)})^{\mathrm{T}}\frac{\partial^2 F_{(k)}^{-i}(\bar\beta_{(k)})}{\partial\beta_{(k)}\partial\beta_{(k)}^{\mathrm{T}}}(\beta_{(k)}-\hat\beta_{(k)})$$

$$\le \phi^{-1}\max_{1\le i\le n}\max_{1\le k\le K}\|\hat\varepsilon_{(k),i}x_{(k),i}\|_2\|\beta_{(k)}-\hat\beta_{(k)}\|_2-\frac{n-1}{2\phi}\|\beta_{(k)}-\hat\beta_{(k)}\|_2^2\min_{1\le i\le n}\min_{1\le k\le K}\lambda_{\min}(\mathcal{H}_k^{-i}(\bar\beta_{(k)})),$$

where $\bar\beta_{(k)}$ lies between $\beta_{(k)}$ and $\hat\beta_{(k)}$. Based on Condition 2, by Boole's inequality and Markov's inequality, for $r\in(0,t_0]$, the following holds:

$$P\left(\max_{1\le i\le n}\max_{1\le k\le K}|\varepsilon_{(k),i}|>Z\right)\le\sum_{i=1}^{n}\sum_{k=1}^{K}P(|\varepsilon_{(k),i}|>Z)\le\sum_{i=1}^{n}\sum_{k=1}^{K}P(\varepsilon_{(k),i}>Z)+\sum_{i=1}^{n}\sum_{k=1}^{K}P(-\varepsilon_{(k),i}>Z)$$

$$\le\sum_{i=1}^{n}\sum_{k=1}^{K}P(e^{r\varepsilon_{(k),i}}>e^{rZ})+\sum_{i=1}^{n}\sum_{k=1}^{K}P(e^{-r\varepsilon_{(k),i}}>e^{rZ})$$

$$\le\sum_{i=1}^{n}\sum_{k=1}^{K}e^{-rZ}(E(e^{r\varepsilon_{(k),i}})+E(e^{-r\varepsilon_{(k),i}}))\le 2nKe^{-rZ}c_0e^{r^2\nu^2/2}.$$

By taking $Z=M\log(nK)$, as $M\to\infty$, we have

$$P\left(\max_{1\le i\le n}\max_{1\le k\le K}|\varepsilon_{(k),i}|>M\log(nK)\right)\le 2c_0e^{-rM\log(nK)+\log(nK)}e^{r^2\nu^2/2}\to 0,$$

which, together with the condition $K/n=o(1)$, implies that $\max_{1\le i\le n}|\varepsilon_{(k),i}|=O_p(\log(n))$ for all $\beta_{(k)}\in\mathcal{B}(\beta_{(k)}^*|\delta^*)$ uniformly in k. Further, using Condition 3, we can derive

$$\max_{1\le i\le n}\max_{1\le k\le K}(F_{(k)}^{-i}(\beta_{(k)})-F_{(k)}^{-i}(\hat\beta_{(k)}))\le\frac{(\log(n))^2\bar p}{n}(\phi^{-1}\delta O_p(1)-C\delta^2(1+o_p(1))). \tag{S.32}$$

By allowing δ to be large enough, the first term $\phi^{-1}\delta O_p(1)$ is dominated by the second term $C\delta^2(1+o_p(1))$. Moreover, it is not difficult to show that $\frac{\partial^2 F_{(k)}^{-i}(\beta_{(k)})}{\partial\beta_{(k)}\partial\beta_{(k)}^{\mathrm{T}}}$ is negative definite for $\beta_{(k)}\in\mathcal{N}_1(\hat\beta_{(k)}|\delta)$ and the sufficiently large n by Condition 3, which implies that the log-likelihood function $F_{(k)}^{-i}(\beta_{(k)})$ is also concave. Further, (S.32) implies that $\max_{1\le i\le n}\max_{1\le k\le K}\|\hat\beta_{(k)}^{-i}-\hat\beta_{(k)}\|_2=O_p(\log(n)\bar p^{1/2}n^{-1})$.

According to the definition of $\tilde\beta_{(k)}^{-i}$, we have $\max_{1\le i\le n}\max_{1\le k\le K}\|\tilde\beta_{(k)}^{-i}-\hat\beta_{(k)}^{-i}\|_2\le B_1+B_2$, where $B_1=\max_{1\le i\le n}\max_{1\le k\le K}\|\hat\beta_{(k)}-\hat\beta_{(k)}^{-i}\|_2$ and $B_2=\max_{1\le i\le n}\max_{1\le k\le K}(n-1)^{-1}(\lambda_{\max}((\mathcal{H}_k^{-i}(\hat\beta_{(k)}))^{-2})\|x_{(k),i}\|_2^2)^{1/2}|\hat\varepsilon_{(k),i}|$. Under Conditions 2 and 3, we can derive that $\max_{1\le i\le n}\max_{1\le k\le K}\|\tilde\beta_{(k)}^{-i}-\hat\beta_{(k)}^{-i}\|_2=O_p(\log(n)\bar p^{1/2}n^{-1})$.

Finally, combining the proofs of (14) and (15), we have

$$\max_{1\le i\le n}\max_{1\le k\le K}\|\hat\beta_{(k)}^{-i}-\beta_{(k)}^*\|_2\le\max_{1\le i\le n}\max_{1\le k\le K}\|\hat\beta_{(k)}^{-i}-\hat\beta_{(k)}\|_2+\max_{1\le k\le K}\|\hat\beta_{(k)}-\beta_{(k)}^*\|_2=O_p(K^{1/2}\bar p^{1/2}n^{-1/2}).$$

This completes the proof of Lemma 1. □

H.3. Proof of Theorem 1

We first focus on the case of $\mathcal{I}_{\mathrm{cor}}=\emptyset$. According to the proof of Ando and Li (2017), to derive (19), it suffices to show that

$$\sup_{w\in\mathcal{W}}|AL(w)/KL(w)-1|\xrightarrow{P}0,\ \sup_{w\in\mathcal{W}}|AP(w)/KL^*(w)|\xrightarrow{P}0,\ \text{and}\ \sup_{w\in\mathcal{W}}|KL(w)/KL^*(w)-1|\xrightarrow{P}0. \tag{S.33}$$

We first establish the first term of (S.33). It is readily shown that $|AL(w) - KL(w)| \le A_{11}(w) + A_{12}(w)$, where $A_{11}(w) = \phi^{-1}|\sum_{i=1}^n b'(\theta_i)(\sum_{k=1}^K w_k\hat{\theta}_{(k),i} - \sum_{k=1}^K w_k\tilde{\theta}_{(k),i})|$ and $A_{12}(w) = \phi^{-1}|\sum_{i=1}^n b(\sum_{k=1}^K w_k\hat{\theta}_{(k),i}) - b(\sum_{k=1}^K w_k\tilde{\theta}_{(k),i})|$. Based on Condition 1, the Jensen's inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}A_{11}(w) &\le \phi^{-1}\left(\sum_{i=1}^n b'(\theta_i)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^n\left(\sum_{k=1}^K w_k(\hat{\theta}_{(k),i} - \tilde{\theta}_{(k),i})\right)^2\right)^{1/2}\\ &\le Cn^{1/2}\left(\sum_{i=1}^n\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2\|\hat{\beta}_{(k)} - \tilde{\beta}_{(k)}^{-i}\|_2^2\right)^{1/2} \le Cn^{1/2}O_p(\log(n)\bar{p}n^{-1/2}) = O_p(\log(n)\bar{p}),\end{aligned}$$

where the last inequality is due to Lemma 1 and Condition 2. Further, by Condition 6 and the third term of (S.33) (the proof is given later),

$$\sup_{w\in\mathcal{W}}\frac{A_{11}(w)}{KL(w)} = O_p\left(\frac{K\bar{p}^{3/2}}{\xi_n}\right)\xrightarrow{P}0. \tag{S.34}$$

Using the Taylor expansion, we have

$$b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}\right) = b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\tilde{\beta}_{(k)}^{-i}\right) + b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\bar{\beta}_{(k)}^{-i}\right)\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\hat{\beta}_{(k)} - \tilde{\beta}_{(k)}^{-i}),$$

where $\bar{\beta}_{(k)}^{-i}$ lies between $\hat{\beta}_{(k)}$ and $\tilde{\beta}_{(k)}^{-i}$. Using Lemma 1, Conditions 1–2, Condition 6 and Cauchy-Schwarz inequality, we deduce that

$$\begin{aligned}\sup_{w\in\mathcal{W}}\frac{A_{12}(w)}{KL(w)} &\le C\xi_n^{-1}\sup_{w\in\mathcal{W}}\left(\sum_{i=1}^n b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\bar{\beta}_{(k)}^{-i}\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^n\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2\|\hat{\beta}_{(k)} - \tilde{\beta}_{(k)}^{-i}\|_2^2\right)^{\frac{1}{2}}\\ &\le C\xi_n^{-1}n^{1/2}O_p(\log(n)\bar{p}n^{-1/2}) = O_p(\xi_n^{-1}\bar{p}\log(n))\xrightarrow{P}0.\end{aligned} \tag{S.35}$$

Combining (S.34) and (S.35), we can derive the first term of (S.33).

Next, we prove the second term of (S.33). Note that

$$|AP(w)| \le \phi^{-1}\left|\sum_{i=1}^n\varepsilon_i\left(\theta_i - \sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*\right)\right| + \phi^{-1}\left|\sum_{i=1}^n\varepsilon_i\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\beta_{(k)}^* - \tilde{\beta}_{(k)}^{-i})\right)\right| \triangleq A_{21}(w) + A_{22}(w).$$

We first prove that $\sup_{w\in\mathcal{W}}A_{21}(w)/KL^*(w)\xrightarrow{P}0$. With Markov's inequality and Conditions 5 and 6, for any $\epsilon > 0$, we have

$$\begin{aligned}&P\left(\sup_{w\in\mathcal{W}}\phi^{-1}\left|\sum_{i=1}^n\varepsilon_i\left(\theta_i - \sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*\right)\right|/KL^*(w) > \epsilon\right) \le P\left(\max_{1\le k\le K}\left|\sum_{i=1}^n\varepsilon_i\left(\theta_i - x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*\right)\right| > \phi\epsilon\xi_n\right)\\ &\le \sum_{k=1}^K CE\left(\left|\sum_{i=1}^n\varepsilon_i\left(\theta_i - x_{(k),i}^{\mathrm{T}}\beta_{(k)}^*\right)\right|^2\right)/(\epsilon\xi_n)^2 \le C\sum_{k=1}^K\|\theta - \theta_{(k)}^*\|_2^2/(\epsilon\xi_n)^2 \le \frac{CKn}{(\epsilon\xi_n)^2}\to 0,\end{aligned}$$

where the third inequality is based on Theorem 2 of Whittle (1960). Thus, $\sup_{w\in\mathcal{W}}A_{21}(w)\ /KL^*(w)\xrightarrow{P}0$. Note that by Condition 4, as $M\to\infty$,

$$P\left(n^{-1/2}\left|\sum_{i=1}^n\varepsilon_i\right| > M\right) = P\left(\left(\sum_{i=1}^n\varepsilon_i\right)^2 > M^2n\right) \le E\left(\sum_{i=1}^n\varepsilon_i\right)^2/(nM^2) = \sum_{i=1}^n E(\varepsilon_i^2)/(nM^2)\to 0,$$

which implies $|\sum_{i=1}^n \varepsilon_i| = O_p(n^{1/2})$. Further, for $A_{22}(w)$,

$$\sup_{w\in\mathcal{W}} \frac{A_{22}(w)}{KL^*(w)} \le C\xi_n^{-1} \max_{1\le i\le n} \max_{1\le k\le K} \|x_{(k),i}^{\mathrm{T}}\|_2 \|\beta_{(k)}^* - \tilde{\beta}_{(k)}^{-i}\|_2 \left|\sum_{i=1}^n \varepsilon_i\right| \le \frac{CK^{1/2}\bar{p}}{\xi_n n^{1/2}} \left|\sum_{i=1}^n \varepsilon_i\right| = O_p(K^{1/2}\bar{p}\xi_n^{-1}) \xrightarrow{P} 0,$$

Hence, the second term of (S.33) holds.

Finally, we give the proof of the third term of (S.33). By Triangle inequality, we have

$$\begin{aligned} |KL(w) - KL^*(w)| \le & \phi^{-1} \left| \sum_{i=1}^n b'(\theta_i) \left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} (\hat{\beta}_{(k)} - \beta_{(k)}^*) \right) \right| \\ & + \phi^{-1} \left| \sum_{i=1}^n \left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right) - b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \beta_{(k)}^* \right) \right) \right| \triangleq A_{31}(w) + A_{32}(w). \end{aligned}$$

Using the Schwarz inequality and Jensen's inequality, based on Lemma 1, Conditions 1–2 and Condition 6,

$$\sup_{w\in\mathcal{W}} \frac{A_{31}(w)}{KL^*(w)} \le C\xi_n^{-1} \sup_{w\in\mathcal{W}} \left(\sum_{i=1}^n b'(\theta_i)^2 \right)^{1/2} \left(\sum_{i=1}^n \sum_{k=1}^K w_k \|x_{(k),i}\|_2^2 \|\hat{\beta}_{(k)} - \beta_{(k)}^*\|_2^2 \right)^{1/2} = O_p(n^{1/2}K^{1/2}\bar{p}\xi_n^{-1}) \xrightarrow{P} 0.$$

Similarly, by Taylor expansion of $b(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)})$ at $\beta_{(k)}^*$, we derive

$$\sup_{w\in\mathcal{W}} \frac{A_{32}(w)}{KL^*(w)} \le \frac{C}{\xi_n} \sup_{w\in\mathcal{W}} \left(\sum_{i=1}^n b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \check{\beta}_{(k)} \right)^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^n \sum_{k=1}^K w_k \|x_{(k),i}\|_2^2 \|\hat{\beta}_{(k)} - \beta_{(k)}^*\|_2^2 \right)^{\frac{1}{2}} = O_p\left(\frac{n^{1/2}K^{1/2}\bar{p}}{\xi_n} \right),$$

where $\check{\beta}_{(k)}$ lies between $\hat{\beta}_{(k)}$ and $\beta_{(k)}^*$. This completes the proof of (19).

In the following, we provide the proof of (20) which can be established if Condition 6 and (S.28) hold according to Lemma S.1. The proof of the second term of (S.28) can refer to that of the second term of (S.33). We focus on the proofs of the first and third terms of (S.28).

First, we provide the proof of the first term of (S.28). It is easy to show that $|AL(w) - R(w)| \le |AL(w) - KL(w)| + |R(w) - KL(w)|$. According to the third term of (S.28) and the proof of (S.33), under the conditions of Theorem 1, we have $\sup_{w\in\mathcal{W}} |AL(w) - KL(w)|/R(w) = O_p(\bar{p}\log(n)/\xi_n) \xrightarrow{P} 0$. Notice that

$$\begin{aligned} |R(w) - KL(w)| \le & \phi^{-1} \left| \sum_{i=1}^n b'(\theta_i) \left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} - \sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} E(\hat{\beta}_{(k)}) \right) \right| \\ & + \phi^{-1} \left| \sum_{i=1}^n \left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right) - E\left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right) \right) \right) \right| \triangleq B_{11}(w) + B_{12}(w). \end{aligned}$$

Further,

$$\begin{aligned} B_{11}(w) \le & \phi^{-1} \left| \sum_{i=1}^n b'(\theta_i) \left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} - \sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \beta_{(k)}^* \right) \right| \\ & + \phi^{-1} \left| \sum_{i=1}^n b'(\theta_i) \left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} \beta_{(k)}^* - \sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}} E(\hat{\beta}_{(k)}) \right) \right| \triangleq B_{111}(w) + B_{112}(w). \end{aligned}$$

Analogous to the proof of $\sup_{w\in\mathcal{W}} A_{31}(w)/KL^*(w) \xrightarrow{P} 0$, we also have

$$\sup_{w\in\mathcal{W}} \frac{B_{111}(w)}{R(w)} = O_p\left(\frac{n^{1/2}K^{1/2}\bar{p}}{\xi_n} \right) \xrightarrow{P} 0. \tag{S.36}$$

We observe that Condition 1 can imply that $\sup_{w\in\mathcal{W}}\sup_{\beta_{(k)}\in\mathcal{B}(\beta^*_{(k)}|\delta^*)} n^{-1}\sum_{i=1}^n E\big(b'\big(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\beta_{(k)}\big)^2\big) \le C < \infty$. Based on Condition 7 and the Cauchy-Schwarz inequality,

$$\sup_{w\in\mathcal{W}}\frac{B_{112}(w)}{R(w)} \le C\xi_n^{-1}\sup_{w\in\mathcal{W}}\left(\sum_{i=1}^n b'(\theta_i)^2\right)^{1/2}\left(\sum_{i=1}^n\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2\|E(\hat\beta_{(k)})-\beta^*_{(k)}\|_2^2\right)^{1/2}$$
$$\le C\xi_n^{-1}\sup_{w\in\mathcal{W}}\left(\sum_{i=1}^n b'(\theta_i)^2\right)^{1/2}\left(\sum_{i=1}^n\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2 E(\|\hat\beta_{(k)}-\beta^*_{(k)}\|_2^2)\right)^{1/2} = O\left(\frac{n^{1/2}K^{1/2}\bar p}{\xi_n}\right)\to 0, \quad \text{(S.37)}$$

where the second inequality is due to Jensen's inequality. For $B_{12}(w)$, applying the Taylor expansion,

$$B_{12}(w) \le \phi^{-1}\left|\sum_{i=1}^n b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\check\beta_{(k)}\right)\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\hat\beta_{(k)}-\beta^*_{(k)})\right|$$
$$+\phi^{-1}\left|\sum_{i=1}^n E\left(b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\check\beta_{(k)}\right)\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\hat\beta_{(k)}-\beta^*_{(k)})\right)\right| \triangleq B_{121}(w)+B_{122}(w),$$

where $\check\beta_{(k)}$ lies between $\hat\beta_{(k)}$ and $\beta^*_{(k)}$. Similar to the proof of $\sup_{w\in\mathcal{W}} A_{32}(w)/KL^*(w) \xrightarrow{P} 0$, we also deduce that

$$\sup_{w\in\mathcal{W}}\frac{B_{121}(w)}{R(w)} = O_p\big(\frac{n^{1/2}K^{1/2}\bar p}{\xi_n}\big)\xrightarrow{P} 0. \quad \text{(S.38)}$$

Employing Condition 7,

$$\sup_{w\in\mathcal{W}}\frac{B_{122}(w)}{R(w)} \le C\xi_n^{-1}\sup_{w\in\mathcal{W}}\sum_{i=1}^n\left(E\left(b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\check\beta_{(k)}\right)^2\right)\right)^{\frac12}\left(E\left(\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2\|\hat\beta_{(k)}-\beta^*_{(k)}\|_2^2\right)\right)^{\frac12}$$
$$= O(n^{1/2}K^{1/2}\bar p\xi_n^{-1})\to 0. \quad \text{(S.39)}$$

Based on the above results, we have $\sup_{w\in\mathcal{W}}|R(w)-KL(w)|/R(w)\xrightarrow{P}0$. Therefore, we finish the proof of the first term of (S.28).

Next, we show the third term of (S.28). Note that

$$|R(w)-KL^*(w)| \le \phi^{-1}\left|\sum_{i=1}^n b'(\theta_i)\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}-\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}E(\hat\beta_{(k)})\right)\right|$$
$$+\phi^{-1}\left|\sum_{i=1}^n\left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\beta^*_{(k)}\right)-E\left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\hat\beta_{(k)}\right)\right)\right)\right| \triangleq B_{21}(w)+B_{22}(w).$$

According to the proofs of (S.37) and (S.39), we also have $\sup_{w\in\mathcal{W}} B_{21}(w)/KL^*(w)\to 0$ and $\sup_{w\in\mathcal{W}} B_{22}(w)/KL^*(w)\to 0$. So the third term of (S.28) holds.

For $\mathcal{I}_{\mathrm{cor}}\neq\emptyset$, we prove the over-consistency of $\tilde w$. Note that $\tilde w = \arg\min_{w\in\mathcal{W}}(AL(w)-AP(w))$. Based on the proof of Theorem 1, we can deduce that

$$\sup_{w\in\mathcal{W}}|AL(w)-AP(w)-KL^*(w)| = \sup_{w\in\mathcal{W}}|AL(w)-KL(w)+KL(w)-KL^*(w)-AP(w)|$$
$$\le \sup_{w\in\mathcal{W}}|AL(w)-KL(w)|+\sup_{w\in\mathcal{W}}|KL(w)-KL^*(w)|+\sup_{w\in\mathcal{W}}|AP(w)|$$
$$= O_p(\bar p\log(n))+O_p(n^{1/2}K^{1/2}\bar p)+O_p(n^{1/2}K^{1/2})+O_p(K^{1/2}\bar p).$$

Thus, $AL(w)-AP(w) = KL^*(w)+O_p(n^{1/2}K^{1/2}\bar p)$. Let $\bar w$ be a weight vector with $\sum_{k\in\mathcal{I}_{\mathrm{cor}}}\bar w_k = 1$. It is easy to see that $KL^*(\bar w)=0$. Therefore, $AL(\bar w)-AP(\bar w) = O_p(n^{1/2}K^{1/2}\bar p)$. Based on the definition of $\tilde w$, we have

$$AL(\tilde w)-AP(\tilde w) = KL^*(\tilde w)+O_p(n^{1/2}K^{1/2}\bar p) \le AL(\bar w)-AP(\bar w) = O_p(n^{1/2}K^{1/2}\bar p).$$

If $\tilde{w}_{\text{cor}} = 1$, then the conclusion holds. Otherwise, if $\tilde{w}_{\text{cor}} \neq 1$, we have

$$\begin{aligned}&(1-\tilde{w}_{\text{cor}})^2 \inf_{w\in\widetilde{\mathcal{W}}}((1-w_{\text{cor}})^{-2}KL^*(w)) + O_p(n^{1/2}K^{1/2}\bar{p})\\ \leq&(1-\tilde{w}_{\text{cor}})^2(1-\tilde{w}_{\text{cor}})^{-2}KL^*(\tilde{w}) + O_p(n^{1/2}K^{1/2}\bar{p}) \leq O_p(n^{1/2}K^{1/2}\bar{p}).\end{aligned}$$

Further, by Condition 6, we can obtain $\tilde{w}_{\text{cor}} \xrightarrow{P} 1$. This completes the proof of Theorem 1. □

H.4. Proof of Corollary 1

Since $y_i = \mu_i + \varepsilon_i$, it is readily shown that

$$\begin{aligned}|CV^*(w) - ACV^*(w)| \leq& \left|\sum_{i=1}^n \phi^{-1}\mu_i\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\hat{\beta}_{(k)}^{-i} - \tilde{\beta}_{(k)}^{-i})\right)\right| + \left|\sum_{i=1}^n \phi^{-1}\varepsilon_i\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}(\hat{\beta}_{(k)}^{-i} - \tilde{\beta}_{(k)}^{-i})\right)\right|\\ &+ \left|\sum_{i=1}^n \phi^{-1}\left(b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}^{-i}\right) - b\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\tilde{\beta}_{(k)}^{-i}\right)\right)\right| \triangleq T_1(w) + T_2(w) + T_3(w).\end{aligned}$$

Referring to the proof of $A_{11}(w)$, we have $\sup_{w\in\mathcal{W}} T_1(w) = O_p(\bar{p}\log(n))$. For $T_2(w)$, under the conditions of Theorem 1, $\sup_{w\in\mathcal{W}} T_2(w) \leq C\max_{1\leq i\leq n}\max_{1\leq k\leq K}\|x_{(k),i}\|_2\|\hat{\beta}_{(k)}^{-i} - \tilde{\beta}_{(k)}^{-i}\|_2\left|\sum_{i=1}^n \varepsilon_i\right| = O_p(\bar{p}\log(n)/n^{1/2})$. By Taylor expansion of $b(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}^{-i})$ at $\tilde{\beta}_{(k)}^{-i}$, similar to the proof of Theorem 1, we have

$$\sup_{w\in\mathcal{W}} T_3(w) \leq \sup_{w\in\mathcal{W}}\left(\sum_{i=1}^n b'\left(\sum_{k=1}^K w_k x_{(k),i}^{\mathrm{T}}\check{\beta}_{(k)}^{-i}\right)^2\right)^{1/2}\left(\sum_{i=1}^n\sum_{k=1}^K w_k\|x_{(k),i}\|_2^2\|\hat{\beta}_{(k)}^{-i} - \tilde{\beta}_{(k)}^{-i}\|_2^2\right)^{1/2} = O_p(\bar{p}\log(n)),$$

where $\check{\beta}_{(k)}^{-i}$ lies between $\hat{\beta}_{(k)}^{-i}$ and $\tilde{\beta}_{(k)}^{-i}$. Thus, $\sup_{w\in\mathcal{W}}|CV^*(w) - ACV^*(w)| = O_p(\bar{p}\log(n))$. Under Condition 6, $\sup_{w\in\mathcal{W}}|ACV^*(w) - CV^*(w)|/\xi_n \xrightarrow{P} 0$. Note that

$$\begin{aligned}&\sup_{w\in\mathcal{W}}\left\{\frac{|ACV^*(w) - CV^*(w)|}{|CV^*(w)|}\right\} \leq \sup_{w\in\mathcal{W}}\left\{\frac{|ACV^*(w) - CV^*(w)|}{KL^*(w)}\right\}\sup_{w\in\mathcal{W}}\left\{\frac{KL^*(w)}{|CV^*(w)|}\right\}\\ =&\sup_{w\in\mathcal{W}}\{|ACV^*(w) - CV^*(w)|/KL^*(w)\}\left[\inf_{w\in\mathcal{W}}\{|CV^*(w)|/KL^*(w)\}\right]^{-1}\\ =&\sup_{w\in\mathcal{W}}\{|ACV^*(w) - CV^*(w)|/KL^*(w)\}\left[1 + \inf_{w\in\mathcal{W}}\{|CV^*(w)|/KL^*(w) - 1\}\right]^{-1}\\ \leq&\sup_{w\in\mathcal{W}}\{|ACV^*(w) - CV^*(w)|/KL^*(w)\}\left[1 + \sup_{w\in\mathcal{W}}\frac{|P(w)|}{KL^*(w)} + \sup_{w\in\mathcal{W}}\left|\frac{L(w)}{KL^*(w)} - 1\right|\right]^{-1}.\end{aligned}$$

According to the proof of Theorem 1, it is readily proved that $\sup_{w\in\mathcal{W}}|P(w)|/KL^*(w) = o_p(1)$ and $\sup_{w\in\mathcal{W}}|L(w)/KL^*(w) - 1| = o_p(1)$. This completes the proof of Corollary 1. □

H.5. Proof of Theorem 2

Denote $\hat{\theta}_{[-i]} = (x_{(1),i}^{\mathrm{T}}\hat{\beta}_{(1)}^{-i}, \ldots, x_{(K),i}^{\mathrm{T}}\hat{\beta}_{(K)}^{-i})^{\mathrm{T}}$. Let $LP(w) = L(w) - P(w)$, where $L(w) = \sum_{i=1}^n \phi^{-1}b'(\theta_i)(\theta_i - w^{\mathrm{T}}\hat{\theta}_{[-i]}) - \phi^{-1}(b(\theta_i) - b(w^{\mathrm{T}}\hat{\theta}_{[-i]}))$ and $P(w) = \sum_{i=1}^n \phi^{-1}\varepsilon_i(w^{\mathrm{T}}\hat{\theta}_{[-i]} - \theta_i)$. Similar to (S.29),

$$-CV(w) = LP(w) - \sum_{i=1}^n(\phi^{-1}\varepsilon_i\theta_i + \phi^{-1}b'(\theta_i)\theta_i - \phi^{-1}b(\theta_i) + c(y_i,\phi)).$$

Thus $\hat{w} = \arg\min_{w\in\mathcal{W}} -CV(w) = \arg\min_{w\in\mathcal{W}} LP(w)$.

Let $r_n = n^{-1/2+\tau}K\bar{p}^2$. Following Li et al. (2022), to prove (21) in Theorem 2, we need to show that there exists a constant Ω such that for the K-dimensional vector $v = (v_1, \ldots, v_K)^{\mathrm{T}}$,

$$\lim_{n\to\infty} P\left(\inf_{\|v\|_2=\Omega, w^0+r_nv\in\mathcal{W}} LP(w^0 + r_nv) > LP(w^0)\right) = 1. \tag{S.40}$$

It is readily shown that

$$\begin{aligned} LP(w^0 + r_nv) - LP(w^0) = &\sum_{i=1}^{n} \phi^{-1} b((w^0 + r_nv)^{\mathrm{T}}\hat{\theta}_{[-i]}) - \phi^{-1} b((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]}) \\ &- \sum_{i=1}^{n} \phi^{-1} b'(\theta_i) r_n v^{\mathrm{T}}\hat{\theta}_{[-i]} - \sum_{i=1}^{n} \phi^{-1}\varepsilon_i r_n v^{\mathrm{T}}\hat{\theta}_{[-i]} \triangleq LP_1 + LP_2 + LP_3. \end{aligned}$$

Based on the Taylor expansion of $b((w^0+r_nv)^{\mathrm{T}}\hat{\theta}_{[-i]})$ at the point w^0, $b((w^0+r_nv)^{\mathrm{T}}\hat{\theta}_{[-i]}) = b((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]}) + b'((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]})r_nv^{\mathrm{T}}\theta_{[-i]} + \frac{1}{2}r_n^2 b''((\bar{w}^0)^{\mathrm{T}}\hat{\theta}_{[-i]})v^{\mathrm{T}}\hat{\theta}_{[-i]}\hat{\theta}_{[-i]}^{\mathrm{T}}v$, where $\bar{w}^0$ lies between w^0 and $w^0 + r_nv$, we have

$$LP_1 = \frac{1}{2}r_n^2\phi^{-1}\sum_{i=1}^{n} b''((\bar{w}^0)^{\mathrm{T}}\hat{\theta}_{[-i]})v^{\mathrm{T}}\hat{\theta}_{[-i]}\hat{\theta}_{[-i]}^{\mathrm{T}}v + \sum_{i=1}^{n}\phi^{-1}b'((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]})r_nv^{\mathrm{T}}\hat{\theta}_{[-i]} \triangleq LP_{11} + LP_{12}.$$

By Condition 9 and the conclusion of Lemma 1, we can deduce that with probability tending to 1, $LP_{11} \geq Cnr_n^2 > 0$.

We next consider $LP_2 + LP_{12}$. Note that $\|\hat{\theta}_{[-i]}\|_2^2 \leq 2(\|\hat{\theta}_{[-i]} - \theta^*_{[i]}\|_2^2 + \|\theta^*_{[i]}\|_2^2)$, where $\theta^*_{[i]} = (x^{\mathrm{T}}_{(1),i}\beta^*_{(1)}, \ldots, x^{\mathrm{T}}_{(K),i}\beta^*_{(K)})^{\mathrm{T}}$. Then, by Condition 2 and Lemma 1, $\sum_{i=1}^n \|\hat{\theta}_{[-i]} - \theta^*_{[i]}\|_2^2 = \sum_{i=1}^n\sum_{k=1}^K (x^{\mathrm{T}}_{(k),i}(\hat{\beta}^{-i}_{(k)} - \beta^*_{(k)}))^2 \leq \sum_{i=1}^n\sum_{k=1}^K \|x_{(k),i}\|_2^2\|\hat{\beta}^{-i}_{(k)} - \beta^*_{(k)}\|_2^2 = O_p(K^2\bar{p}^2)$. Using Condition 8, $\sum_{i=1}^n \|\theta^*_{[i]}\|_2^2 \leq \sum_{i=1}^n\sum_{k=1}^K \|x_{(k),i}\|_2^2\|\beta^*_{(k)}\|_2^2 \leq CnK\bar{p}^2$. Thus, $\sum_{i=1}^n \|\hat{\theta}_{[-i]}\|_2^2 = O_p(nK\bar{p}^2)$. Also, by c_r inequality, $\|\hat{\theta}_{[-i]}\|_2^4 \leq 8(\|\hat{\theta}_{[-i]} - \theta^*_{[i]}\|_2^4 + \|\theta^*_{[i]}\|_2^4)$. Similarly, $\sum_{i=1}^n \|\hat{\theta}_{[-i]} - \theta^*_{[i]}\|_2^4 \leq \sum_{i=1}^n(\sum_{k=1}^K \|x_{(k),i}\|_2^2\|\hat{\beta}^{-i}_{(k)} - \beta^*_{(k)}\|_2^2)^2 = O_p(n^{-1}K^4\bar{p}^4)$. And $\sum_{i=1}^n \|\theta^*_{[i]}\|_2^4 \leq \sum_{i=1}^n(\sum_{k=1}^K \|x_{(k),i}\|_2^2\|\beta^*_{(k)}\|_2^2)^2 \leq CnK^2\bar{p}^4$. Hence, $\sum_{i=1}^n \|\hat{\theta}_{[-i]}\|_2^4 = O_p(nK^2\bar{p}^4)$.

Furthermore,

$$\begin{aligned} LP_2 + LP_{12} = &\sum_{i=1}^{n}\phi^{-1}r_n(b'((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]}) - b'((w^0)^{\mathrm{T}}\hat{\theta}_{[i]}))v^{\mathrm{T}}\hat{\theta}_{[-i]} \\ &+ \sum_{i=1}^{n}\phi^{-1}r_n(b'((w^0)^{\mathrm{T}}\hat{\theta}_{[i]}) - b'(\theta_i))v^{\mathrm{T}}\hat{\theta}_{[-i]} \triangleq LP_{21} + LP_{22}, \end{aligned}$$

where $\hat{\theta}_{[i]} = (x^{\mathrm{T}}_{(1),i}\hat{\beta}_{(1)}, \ldots, x^{\mathrm{T}}_{(K),i}\hat{\beta}_{(K)})^{\mathrm{T}}$. Based on the Taylor expression, $b'((w^0)^{\mathrm{T}}\hat{\theta}_{[-i]}) = b'(\sum_{k=1}^K w_k^0 x^{\mathrm{T}}_{(k),i}\hat{\beta}_{(k)}) + b''(\sum_{k=1}^K w_k^0 x^{\mathrm{T}}_{(k),i}\hat{\tilde{\beta}}^{-i}_{(k)})\sum_{k=1}^K w_k^0 x^{\mathrm{T}}_{(k),i}(\hat{\beta}^{-i}_{(k)} - \hat{\beta}_{(k)})$, where $\hat{\tilde{\beta}}^{-i}_{(k)}$ lies between $\hat{\beta}^{-i}_{(k)}$ and $\hat{\beta}_{(k)}$. Then,

$$\begin{aligned} LP_{21} \leq &\phi^{-1}r_n\left(\sum_{i=1}^{n} b''\left(\sum_{k=1}^{K} w_k^0 x^{\mathrm{T}}_{(k),i}\hat{\tilde{\beta}}^{-i}_{(k)}\right)^2\right)^{1/2}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} w_k^0\|x_{(k),i}\|_2^2\|v\|_2^2\|\hat{\theta}_{[-i]}\|_2^2\|\hat{\beta}^{-i}_{(k)} - \hat{\beta}_{(k)}\|_2^2\right)^{1/2} \\ \leq &C\Omega r_n n^{1/2}O_p((\bar{p}^4 n^{-1}K)^{1/2}\log(n)) = O_p(r_nK^{1/2}\bar{p}^2\log(n)). \end{aligned}$$

As $\psi_n = E(\|\breve{\mu} - \breve{\mu}(w^0)\|_2^2)$, then $\|\breve{\mu} - \breve{\mu}(w^0)\|_2 = O_p(\psi_n^{1/2})$. For LP_{22}, by Condition 1 and the Cauchy-Schwarz inequality, we have

$$LP_{22} = \sum_{i=1}^{n} \phi^{-1} r_n b'' \left(\sum_{k=1}^{K} \hat{w}_k^* x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right)^{-1/2} (b'((w^0)^{\mathrm{T}} \hat{\theta}_{[i]}) - b'(\theta_i)) b'' \left(\sum_{k=1}^{K} \hat{w}_k^* x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right)^{1/2} v^{\mathrm{T}} \hat{\theta}_{[-i]}$$
$$\leq C r_n \|\breve{\mu} - \breve{\mu}(w^0)\|_2 \left(\sum_{i=1}^{n} b'' \left(\sum_{k=1}^{K} \hat{w}_k^* x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right) \|v\|_2^2 \|\hat{\theta}_{[-i]}\|_2^2 \right)^{1/2}$$
$$\leq C r_n \|\breve{\mu} - \breve{\mu}(w^0)\|_2 \left(\left(\sum_{i=1}^{n} b'' \left(\sum_{k=1}^{K} \hat{w}_k^* x_{(k),i}^{\mathrm{T}} \hat{\beta}_{(k)} \right)^2 \right)^{1/2} \left(\sum_{i=1}^{n} \|\hat{\theta}_{[-i]}\|_2^4 \right)^{1/2} \right)^{1/2} = O_p(r_n \psi_n^{1/2} n^{1/2} K^{1/2} \bar{p}).$$

We turn to consider LP_3. Note that $LP_3 \leq |\sum_{i=1}^{n} \phi^{-1} r_n \varepsilon_i v^{\mathrm{T}} \theta_{[i]}^*| + |\sum_{i=1}^{n} \phi^{-1} r_n \varepsilon_i v^{\mathrm{T}} (\hat{\theta}_{[-i]} - \theta_{[i]}^*)| \triangleq LP_{31} + LP_{32}$. We first show that $LP_{31} = O_p(r_n n^{1/2} K^{1/2} \bar{p})$. Based on Condition 4, using Theorem 2 of Whittle (1960), as $M \to \infty$,

$$P\left(\left| \sum_{i=1}^{n} \phi^{-1} r_n \varepsilon_i v^{\mathrm{T}} \theta_{[i]}^* \right| > M r_n n^{1/2} K^{1/2} \bar{p} \right) \leq E \left| \sum_{i=1}^{n} \phi^{-1} r_n \varepsilon_i v^{\mathrm{T}} \theta_{[i]}^* \right|^2 / (M r_n n^{1/2} K^{1/2} \bar{p})^2$$
$$\leq C \sum_{i=1}^{n} r_n^2 \|v\|_2^2 \|\theta_{[i]}^*\|_2^2 / (M r_n n^{1/2} K^{1/2} \bar{p})^2 \leq C r_n^2 n K \bar{p}^2 / (M r_n n^{1/2} K^{1/2} \bar{p})^2 \to 0.$$

On the other hand, $LP_{32} \leq C r_n (\sum_{i=1}^{n} \varepsilon_i^2)^{1/2} (\sum_{i=1}^{n} \|v\|_2^2 \|\hat{\theta}_{[-i]} - \theta_{[i]}^*\|_2^2)^{1/2} = O_p(r_n n^{1/2} K \bar{p})$. To sum up, by Condition 10, we can see that LP_{11} dominates LP_{12}, LP_2 and LP_3 asymptotically. Thus, (S.40) holds. This completes the proof of (21).

Next, we prove (22) of Theorem 2. Similar to the proof of (21), we also illustrate that there exists a constant Ω such that for the K-dimensional vector $v = (v_1, \ldots, v_K)^{\mathrm{T}}$,

$$\lim_{n \to \infty} P\left(\inf_{\|v\|_2 = \Omega, w^0 + r_n v \in \mathcal{W}} ALP(w^0 + r_n v) > ALP(w^0) \right) = 1, \tag{S.41}$$

where $ALP(w) = AL(w) - AP(w)$, and $AL(w)$ and $AP(w)$ have been defined in the proof of Theorem 1. It is noteworthy that

$$ALP(w^0 + r_n v) - ALP(w^0) = \sum_{i=1}^{n} \phi^{-1} b((w^0 + r_n v)^{\mathrm{T}} \tilde{\theta}_{[-i]}) - \phi^{-1} b((w^0)^{\mathrm{T}} \tilde{\theta}_{[-i]})$$
$$- \sum_{i=1}^{n} \phi^{-1} b'(\theta_i) r_n v^{\mathrm{T}} \tilde{\theta}_{[-i]} - \sum_{i=1}^{n} \phi^{-1} \varepsilon_i r_n v^{\mathrm{T}} \tilde{\theta}_{[-i]} \triangleq ALP_1 + ALP_2 + ALP_3,$$

where $\tilde{\theta}_{[-i]} = (x_{(1),i}^{\mathrm{T}} \tilde{\beta}_{(1)}^{-i}, \ldots, x_{(K),i}^{\mathrm{T}} \tilde{\beta}_{(K)}^{-i})^{\mathrm{T}}$. Based on the Taylor expansion, $ALP_1 = ALP_{11} + ALP_{12}$, where $ALP_{11} = \frac{1}{2} r_n^2 \phi^{-1} \sum_{i=1}^{n} b''((\bar{w}^0)^{\mathrm{T}} \tilde{\theta}_{[-i]}) v^{\mathrm{T}} \tilde{\theta}_{[-i]} \tilde{\theta}_{[-i]}^{\mathrm{T}} v$ and $ALP_{12} = \sum_{i=1}^{n} \phi^{-1} b'((w^0)^{\mathrm{T}} \tilde{\theta}_{[-i]}) r_n v^{\mathrm{T}} \tilde{\theta}_{[-i]}$. Based on Condition 9 and the conclusion of Lemma 1, we have $ALP_{11} \geq C n r_n^2 > 0$ with probability tending to 1.

We next turn to consider $ALP_2 + ALP_{12}$. Similar to the proof of $\sum_{i=1}^{n} \|\hat{\theta}_{[-i]}\|_2^2 = O_p(nK\bar{p}^2)$ and $\sum_{i=1}^{n} \|\hat{\theta}_{[-i]}\|_2^4 = O_p(nK^2\bar{p}^4)$, by Condition 2 and Lemma 1, we also have $\sum_{i=1}^{n} \|\tilde{\theta}_{[-i]}\|_2^2 = O_p(nK\bar{p}^2)$ and $\sum_{i=1}^{n} \|\tilde{\theta}_{[-i]}\|_2^4 = O_p(nK^2\bar{p}^4)$. Further,

$$ALP_2 + ALP_{12} = \sum_{i=1}^{n} \phi^{-1} r_n (b'((w^0)^{\mathrm{T}} \tilde{\theta}_{[-i]}) - b'((w^0)^{\mathrm{T}} \hat{\theta}_{[i]})) v^{\mathrm{T}} \tilde{\theta}_{[-i]}$$
$$+ \sum_{i=1}^{n} \phi^{-1} r_n (b'((w^0)^{\mathrm{T}} \hat{\theta}_{[i]}) - b'(\theta_i)) v^{\mathrm{T}} \tilde{\theta}_{[-i]} \triangleq ALP_{21} + ALP_{22},$$

where $\hat{\theta}_{[i]} = (x_{(1),i}^{\mathrm{T}}\hat{\beta}_{(1)}, \ldots, x_{(K),i}^{\mathrm{T}}\hat{\beta}_{(K)})^{\mathrm{T}}$. Based on the Taylor expression, $b'((w^0)^{\mathrm{T}}\tilde{\theta}_{[-i]}) = b'(\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}) + b''(\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}\tilde{\tilde{\beta}}_{(k)}^{-i})\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}(\tilde{\beta}_{(k)}^{-i} - \hat{\beta}_{(k)})$, where $\tilde{\tilde{\beta}}_{(k)}^{-i}$ lies between $\tilde{\beta}_{(k)}^{-i}$ and $\hat{\beta}_{(k)}$. Then,

$$ALP_{21} = \sum_{i=1}^n \phi^{-1} r_n b''\left(\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}\tilde{\tilde{\beta}}_{(k)}^{-i}\right)\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}(\tilde{\beta}_{(k)}^{-i} - \hat{\beta}_{(k)})v^{\mathrm{T}}\tilde{\theta}_{[-i]}$$
$$\le \phi^{-1} r_n\left(\sum_{i=1}^n b''\left(\sum_{k=1}^K w_k^0 x_{(k),i}^{\mathrm{T}}\tilde{\tilde{\beta}}_{(k)}^{-i}\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^n\sum_{k=1}^K w_k^0\|x_{(k),i}\|_2^2\|v\|_2^2\|\tilde{\theta}_{[-i]}\|_2^2\|\tilde{\beta}_{(k)}^{-i} - \hat{\beta}_{(k)}\|_2^2\right)^{\frac{1}{2}} = O_p(r_n K^{1/2}\bar{p}^2\log(n)).$$

For ALP_{22}, by Condition 1 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}ALP_{22} &= \sum_{i=1}^n \phi^{-1} r_n b''\left(\sum_{k=1}^K \hat{w}_k^* x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}\right)^{-1/2}(b'((w^0)^{\mathrm{T}}\hat{\theta}_{[i]}) - b'(\theta_i))b''\left(\sum_{k=1}^K \hat{w}_k^* x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}\right)^{1/2} v^{\mathrm{T}}\tilde{\theta}_{[-i]}\\
&\le Cr_n\|\check{\mu} - \check{\mu}(w^0)\|_2\left(\sum_{i=1}^n b''\left(\sum_{k=1}^K \hat{w}_k^* x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}\right)\|v\|_2^2\|\tilde{\theta}_{[-i]}\|_2^2\right)^{1/2}\\
&\le Cr_n\|\check{\mu} - \check{\mu}(w^0)\|_2\left(\left(\sum_{i=1}^n b''\left(\sum_{k=1}^K \hat{w}_k^* x_{(k),i}^{\mathrm{T}}\hat{\beta}_{(k)}\right)^2\right)^{1/2}\left(\sum_{i=1}^n\|\tilde{\theta}_{[-i]}\|_2^4\right)^{1/2}\right)^{1/2} = O_p(r_n\psi_n^{1/2}n^{1/2}K^{1/2}\bar{p}).\end{aligned}$$

We next consider ALP_3. Notice that $ALP_3 \le |\sum_{i=1}^n \phi^{-1} r_n\varepsilon_i v^{\mathrm{T}}\theta_{[i]}^*| + |\sum_{i=1}^n \phi^{-1} r_n\varepsilon_i v^{\mathrm{T}}(\tilde{\theta}_{[-i]} - \theta_{[i]}^*)| \triangleq ALP_{31} + ALP_{32}$. It is known that $ALP_{31} = LP_{31} = O_p(r_n n^{1/2}K^{1/2}\bar{p})$, and $ALP_{32} \le Cr_n(\sum_{i=1}^n\varepsilon_i^2)^{1/2}(\sum_{i=1}^n\|v\|_2^2\|\tilde{\theta}_{[-i]} - \theta_{[i]}^*\|_2^2)^{1/2} = O_p(r_n n^{1/2}K\bar{p})$. By Condition 10, it is readily shown that ALP_{11} dominates ALP_{12}, ALP_2 and ALP_3 asymptotically. Thus, this completes the proof of Theorem 2. □

H.6. Proof of Corollary 2

Based on the conclusion of Theorem 2, we can derive that

$$\|\hat{w} - \tilde{w}\|_2 \le \|\hat{w} - w^0\|_2 + \|\tilde{w} - w^0\|_2 = O_p(n^{-1/2+\tau}K\bar{p}^2).$$

This completes the proof of Corollary 2. □

H.7. Proof of Theorem 3

Without loss of generality, suppose that the k_0th model is a correct model. Then, based on Lemma 1, we can derive that $\|\Pi_{(k_0)}\hat{\beta}_{(k_0)} - \beta\|_2 = O_p(n^{-1/2}p^{1/2})$. Further, by Lemma 1, we also have $\max_{1\le i\le n}\|\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \Pi_{(k_0)}\hat{\beta}_{(k_0)}\|_2 = O_p(n^{-1}p^{1/2}\log(n))$. Denote by $w_{k_0}^0$ the weight vector whose k_0th element is 1 and others are 0. Thus, $ACV(\tilde{w}) - ACV(w_{k_0}^0) \ge 0$.

We first focus on $ACV(w_{k_0}^0)$. For $ACV(w_{k_0}^0)$, we can derive that $ACV(w_{k_0}^0) = \sum_{i=1}^n(\phi^{-1}y_i x_i^{\mathrm{T}}\beta - \phi^{-1}b(x_i^{\mathrm{T}}\beta) + c(y_i,\phi)) + \sum_{i=1}^n(\phi^{-1}y_i x_i^{\mathrm{T}}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta) + \phi^{-1}(b(x_i^{\mathrm{T}}\beta) - b(x_i^{\mathrm{T}}\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i})))$. Based on the Taylor expansion, we have $b(x_i^{\mathrm{T}}\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i}) = b(x_i^{\mathrm{T}}\beta) + b'(x_i^{\mathrm{T}}\beta)x_i^{\mathrm{T}}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta) + 1/2(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta)^{\mathrm{T}}b''(x_i^{\mathrm{T}}\bar{\beta}_{(k_0)})x_i x_i^{\mathrm{T}}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta)$, where $\bar{\beta}_{(k_0)}$ lies between β and $\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i}$. Thus,

$$\begin{aligned}ACV(w_{k_0}^0) &= \sum_{i=1}^n(\phi^{-1}y_i x_i^{\mathrm{T}}\beta - \phi^{-1}b(x_i^{\mathrm{T}}\beta) + c(y_i,\phi)) + \sum_{i=1}^n\phi^{-1}\varepsilon_i x_i^{\mathrm{T}}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta)\\
&\quad - \sum_{i=1}^n\frac{1}{2}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta)^{\mathrm{T}}b''(x_i^{\mathrm{T}}\bar{\beta}_{(k_0)})x_i x_i^{\mathrm{T}}(\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i} - \beta) \triangleq ACV_1^0 + ACV_2^0 + ACV_3^0.\end{aligned}$$

Further, denote $\tilde{\beta}^{-i}(\tilde{w}) = \sum_{k=1}^{K} \tilde{w}_k \Pi_{(k)} \tilde{\beta}_{(k)}^{-i}$. Then,

$$ACV(w_{k_0}^0) - ACV(\tilde{w}) = \sum_{i=1}^{n} \phi^{-1} y_i x_i^{\mathrm{T}} (\beta - \hat{\beta}(\tilde{w})) + \sum_{i=1}^{n} \phi^{-1} y_i x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w}))$$
$$+ \sum_{i=1}^{n} \phi^{-1} (b(x_i^{\mathrm{T}} \tilde{\beta}^{-i}(\tilde{w})) - b(x_i^{\mathrm{T}} \hat{\beta}(\tilde{w}))) + \sum_{i=1}^{n} \phi^{-1} (b(x_i^{\mathrm{T}} \hat{\beta}(\tilde{w})) - b(x_i^{\mathrm{T}} \beta)) + ACV_2^0 + ACV_3^0. \tag{S.42}$$

We next discuss the magnitude of each term in $ACV(w_{k_0}^0) - ACV(\tilde{w})$. Note that $E(\| \sum_{i=1}^{n} \varepsilon_i x_i \|_2^2) = E(\varepsilon^{\mathrm{T}} X X^{\mathrm{T}} \varepsilon) = tr(X X^{\mathrm{T}} \mathrm{Cov}(\varepsilon)) = \sum_{i=1}^{n} \|x_i\|_2^2 \mathrm{Vay}(y_i) \le Cnp$. Thus, by Markov's inequality, we have $\| \sum_{i=1}^{n} \varepsilon_i x_i \|_2 = O_p(\sqrt{np})$. Additionally, $\| \sum_{i=1}^{n} \mu_i x_i \|_2 \le \sum_{i=1}^{n} |\mu_i| \|x_i\|_2 \le (\sum_{i=1}^{n} \mu_i^2)^{1/2} (\sum_{i=1}^{n} \|x_i\|_2^2)^{1/2} \le Cnp^{1/2}$. Then, $\| \sum_{i=1}^{n} y_i x_i \|_2 \le \| \sum_{i=1}^{n} \mu_i x_i \|_2 + \| \sum_{i=1}^{n} \varepsilon_i x_i \|_2 = O_p(np^{1/2})$. We first consider the second term in $ACV(w_{k_0}^0) - ACV(\tilde{w})$. By Lemma 1, $\max_{1 \le i \le n} \|\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w})\|_2 \le \max_{1 \le i \le n} \max_{1 \le k \le K} \|\hat{\beta}_{(k)} - \tilde{\beta}_{(k)}^{-i}\|_2 = O_p(n^{-1} \log(n) p^{1/2})$. So

$$\Big| \sum_{i=1}^{n} \phi^{-1} y_i x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w})) \Big| \le C \sum_{i=1}^{n} |y_i| \|x_i\|_2 \max_{1 \le i \le n} \|\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w})\|_2 = O_p(\log(n) p), \tag{S.43}$$

where the last equality is based on the fact that $\sum_{i=1}^{n} y_i^2 = O_p(n)$ since $E(\sum_{i=1}^{n} y_i^2) \le \sum_{i=1}^{n} (\mathrm{Var}(y_i) + (E(y_i))^2) \le Cn$ based on Condition 1. By Taylor expansion, we deduce that

$$\sum_{i=1}^{n} \phi^{-1} (b(x_i^{\mathrm{T}} \tilde{\beta}^{-i}(\tilde{w})) - b(x_i^{\mathrm{T}} \hat{\beta}(\tilde{w}))) = \sum_{i=1}^{n} \phi^{-1} b'(x_i^{\mathrm{T}} \bar{\beta}^{-i}(\tilde{w})) x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w}))$$
$$\le C \Bigg(\sum_{i=1}^{n} (b'(x_i^{\mathrm{T}} \bar{\beta}^{-i}(\tilde{w})))^2 \Bigg)^{1/2} \Bigg(\sum_{i=1}^{n} \|x_i\|_2^2 \Bigg)^{1/2} \max_{1 \le i \le n} \|\hat{\beta}(\tilde{w}) - \tilde{\beta}^{-i}(\tilde{w})\|_2 = O_p(p \log(n)), \tag{S.44}$$

where $\bar{\beta}^{-i}(\tilde{w})$ lies between $\hat{\beta}(\tilde{w})$ and $\tilde{\beta}^{-i}(\tilde{w})$, and the last equality is implied by Conditions 1 and 2.

Notice that $b(x_i^{\mathrm{T}} \hat{\beta}(\tilde{w})) = b(x_i^{\mathrm{T}} \beta) + b'(x_i^{\mathrm{T}} \beta) x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \beta) + 1/2 b''(x_i^{\mathrm{T}} \bar{\beta}(\tilde{w})) (\hat{\beta}(\tilde{w}) - \beta)^{\mathrm{T}} x_i x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \beta)$ where $\bar{\beta}(\tilde{w})$ lies between $\hat{\beta}(\tilde{w})$ and β. Then, for the fourth term in $ACV(w_{k_0}^0) - ACV(\tilde{w})$, we have

$$\sum_{i=1}^{n} \phi^{-1} (b(x_i^{\mathrm{T}} \hat{\beta}(\tilde{w})) - b(x_i^{\mathrm{T}} \beta))$$
$$= \sum_{i=1}^{n} \phi^{-1} b'(x_i^{\mathrm{T}} \beta) x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \beta) + \sum_{i=1}^{n} \phi^{-1} \frac{1}{2} b''(x_i^{\mathrm{T}} \bar{\beta}(\tilde{w})) (\hat{\beta}(\tilde{w}) - \beta)^{\mathrm{T}} x_i x_i^{\mathrm{T}} (\hat{\beta}(\tilde{w}) - \beta) \triangleq S_1 + S_2. \tag{S.45}$$

Combining with the first term in $ACV(w_{k_0}^0) - ACV(\tilde{w})$ and S_1, using $\| \sum_{i=1}^{n} \varepsilon_i x_i \|_2 = O_p(\sqrt{np})$,

$$\left| \sum_{i=1}^{n} \phi^{-1} y_i x_i^{\mathrm{T}} (\beta - \hat{\beta}(\tilde{w})) + S_1 \right| = \left| \sum_{i=1}^{n} \phi^{-1} \varepsilon_i x_i^{\mathrm{T}} (\beta - \hat{\beta}(\tilde{w})) \right| = \|\beta - \hat{\beta}(\tilde{w})\|_2 O_p(n^{1/2} p^{1/2}). \tag{S.46}$$

Furthermore, by Condition 11, $S_2 \ge Cn \|\beta - \hat{\beta}(\tilde{w})\|_2^2 (1 + o_p(1))$. For ACV_2^0, we can derive that

$$|ACV_2^0| \le C \sum_{i=1}^{n} |\varepsilon_i| \|x_i\|_2 \|\tilde{\beta}_{(k_0)}^{-i} - \hat{\beta}_{(k_0)}\|_2 + C \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2 \|\Pi_{(k_0)} \hat{\beta}_{(k_0)} - \beta\|_2 = O_p(p \log(n)). \tag{S.47}$$

For ACV_3^0, we can see that

$$|ACV_3^0| \le C\sum_{i=1}^{n}|b''(x_i^{\mathrm{T}}\bar{\beta}_{(k_0)})|\|x_i\|_2^2\|\Pi_{(k_0)}\tilde{\beta}_{(k_0)}^{-i}-\beta\|_2^2 = O_p\big(np\times\frac{p}{n}\big) = O_p(p^2). \tag{S.48}$$

Denote $u=\hat{\beta}(\tilde{w})-\beta$. Combining with (S.43)–(S.48), we have

$$\begin{aligned}&ACV(w_{k_0}^0)-ACV(\tilde{w}) \ge Cn(1+o_p(1))\|u\|_2^2 - \|u\|_2 O_p(n^{1/2}p^{1/2}) - O_p(p\log(n)) - O_p(p^2)\\ &= p^2\log(n)(C(1+o_p(1))\|u\|_2^2 np^{-2}/\log(n) - \|u\|_2 O_p(n^{1/2}p^{-3/2}/\log(n)) - O_p(1)) \triangleq S_3.\end{aligned}$$

Based on the definition of $\tilde{w}$, we deduce that $P(ACV(w_{k_0}^0)-ACV(\tilde{w})\ge 0)=0$, which shows that S_3 is negative with probability tending to one. Hence, $n(\log(n))^{-1}p^{-2}\|u\|_2^2$ has to be bounded with probability tending to one, that is, $n(\log(n))^{-1}p^{-2}\|\hat{\beta}(\tilde{w})-\beta\|_2^2 = O_p(1)$. □

H.8. Proof of Theorem 4

To derive (23) and (24) in Theorem 4, it is sufficient to show that for any $\epsilon>0$,

$$P\left(\left|\frac{\inf_{w\in\mathcal{W}}KL(w)}{KL(\tilde{w}^s)}-1\right|>\epsilon\right)\to 0 \text{ and } P\left(\left|\frac{\inf_{w\in\mathcal{W}}R(w)}{R(\tilde{w}^s)}-1\right|>\epsilon\right)\to 0, \tag{S.49}$$

respectively. Let $v_n = KL(w^A)-\inf_{w\in\mathcal{W}}KL(w)$. Recall that $\arg\max_{w\in W}ACV(w)$ is equivalent to $\arg\min_{w\in W}ALP(w)$ with $ALP(w)=AL(w)-AP(w)$. Thus, $\tilde{w}^s=\arg\min_{w\in\mathcal{W}_S}ALP(w)$.

We first prove the first part of (S.49). Note that

$$\begin{aligned}&\left|\frac{\inf_{w\in\mathcal{W}}KL(w)}{KL(\tilde{w}^s)}-1\right| = \left|\frac{KL(\tilde{w}^s)-\inf_{w\in\mathcal{W}}KL(w)}{KL(\tilde{w}^s)}\right| = \left|\frac{KL(\tilde{w}^s)-KL(w^A)+v_n}{KL(\tilde{w}^s)}\right|\\ \le&\left|\frac{KL(\tilde{w}^s)-AL(\tilde{w}^s)+AP(\tilde{w}^s)+v_n+AL(w^A)-AP(w^A)-KL(w^A)}{KL(\tilde{w}^s)}\right|\\ \le&\sup_{w\in\mathcal{W}}\left|\frac{KL(w)-AL(w)}{KL(w)}\right|+\left|\frac{v_n}{KL(\tilde{w}^s)}\right|+\left|\frac{KL(w^A)-AL(w^A)}{KL(\tilde{w}^s)}\right|+\left|\frac{AP(\tilde{w}^s)}{KL(\tilde{w}^s)}\right|+\left|\frac{AP(w^A)}{KL(\tilde{w}^s)}\right|\\ \triangleq&L_1+L_2+L_3+L_4+L_5.\end{aligned}$$

For the first term L_1, according to (S.33) in the proof of Theorem 1, we have $L_1\xrightarrow{P}0$. For the second term L_2, we have $L_2\le\left|\frac{v_n}{\inf_{w\in\mathcal{W}}KL(w)}\right|\le\left|\frac{v_n}{\inf_{w\in\mathcal{W}}KL^*(w)}\right|\times\left|\frac{\inf_{w\in\mathcal{W}}KL^*(w)}{\inf_{w\in\mathcal{W}}KL(w)}\right|$. By the ALL property, $|v_n/\inf_{w\in\mathcal{W}}KL^*(w)|\xrightarrow{P}0$. Further, using the third term of (S.33) and the ALL property,

$$\begin{aligned}&\left|\frac{\inf_{w\in\mathcal{W}}KL^*(w)}{\inf_{w\in\mathcal{W}}KL(w)}\right|\le\left|\frac{KL^*(w^A)}{KL(w^A)-v_n}\right|\le\sup_{w\in\mathcal{W}}\left|\frac{KL^*(w)}{KL(w)-v_n}\right| = \left\{\inf_{w\in\mathcal{W}}\left|\frac{KL(w)-v_n}{KL^*(w)}\right|\right\}^{-1}\\ \le&\left\{1-\sup_{w\in\mathcal{W}}\left|\frac{KL(w)-KL^*(w)}{KL^*(w)}\right|-\sup_{w\in\mathcal{W}}\left|\frac{v_n}{KL^*(w)}\right|\right\}^{-1}\xrightarrow{P}1.\end{aligned}$$

Thus, we have $L_2\xrightarrow{P}0$. Next, we turn to consider L_3, that is, $L_3\le\sup_{w\in\mathcal{W}}\left|\frac{KL(w)-AL(w)}{KL(w)}\right|\times\left|\frac{KL(w^A)}{KL(\tilde{w}^s)}\right|$. By (S.33), we have $\sup_{w\in\mathcal{W}}|(KL(w)-AL(w))/KL(w)|\xrightarrow{P}0$. Also, based on $L_2\xrightarrow{P}0$, we have

$$\left|\frac{KL(w^A)}{KL(\tilde{w}^s)}\right| = \left|\frac{\inf_{w\in\mathcal{W}}KL(w)+v_n}{KL(\tilde{w}^s)}\right| = \left|\frac{\inf_{w\in\mathcal{W}}KL(w)}{KL(\tilde{w}^s)}\right|+\left|\frac{v_n}{KL(\tilde{w}^s)}\right| < 1+o_p(1). \tag{S.50}$$

Therefore, we can derive $L_3 \xrightarrow{P} 0$. For the term L_4, we deduce that $L_4 \leq \sup_{w\in\mathcal{W}}\left|\frac{AP(w)}{KL(w)}\right| \leq \sup_{w\in\mathcal{W}}\left|\frac{AP(w)}{KL^*(w)}\right|\sup_{w\in\mathcal{W}}\left|\frac{KL^*(w)}{KL(w)}\right|$. Based on the second term of (S.33), we have $\sup_{w\in\mathcal{W}}|AP(w)/KL^*(w)| \xrightarrow{P} 0$. In addition, by the third term of (S.33),

$$\sup_{w\in\mathcal{W}}\left|\frac{KL^*(w)}{KL(w)}\right| = \left\{\inf_{w\in\mathcal{W}}\left|\frac{KL(w)}{KL^*(w)}\right|\right\}^{-1} \leq \left\{1-\sup_{w\in\mathcal{W}}\left|\frac{KL(w)-KL^*(w)}{KL^*(w)}\right|\right\}^{-1} \xrightarrow{P} 1.$$

Based on the above derivation, we derive $L_4 \xrightarrow{P} 0$. Further, based on the proof of $L_4 \xrightarrow{P} 0$ and $\left|\frac{KL(w^A)}{KL(\tilde{w}^s)}\right| < 1+o_p(1)$, we have $L_5 \leq \sup_{w\in\mathcal{W}}\left|\frac{AP(w)}{KL(w)}\right| \times \left|\frac{KL(w^A)}{KL(\tilde{w}^s)}\right| \xrightarrow{P} 0$. At this point, we show that (23) holds.

Next, we give the proof of (24). Notice that

$$\begin{aligned}
&\left|\frac{\inf_{w\in\mathcal{W}}R(w)}{R(\tilde{w}^s)}-1\right| = \left|\frac{R(\tilde{w}^s)-\inf_{w\in\mathcal{W}}R(w)}{R(\tilde{w}^s)}\right| \\
\leq&\left|\frac{R(\tilde{w}^s)-AL(\tilde{w}^s)+AP(\tilde{w}^s)+ALP(w^A)-\inf_{w\in\mathcal{W}}KL(w)+\inf_{w\in\mathcal{W}}KL(w)-\inf_{w\in\mathcal{W}}R(w)}{R(\tilde{w}^s)}\right| \\
\leq& \sup_{w\in\mathcal{W}}\left|\frac{R(w)-AL(w)}{R(w)}\right| + \left|\frac{AP(\tilde{w}^s)}{R(\tilde{w}^s)}\right| + \left|\frac{ALP(w^A)-\inf_{w\in\mathcal{W}}KL(w)}{R(\tilde{w}^s)}\right| + \left|\frac{\inf_{w\in\mathcal{W}}KL(w)-\inf_{w\in\mathcal{W}}R(w)}{R(\tilde{w}^s)}\right| \\
\triangleq& R_1+R_2+R_3+R_4.
\end{aligned}$$

According to (S.28), it is readily shown that $R_1 \xrightarrow{P} 0$. For the second term R_2, based on the second term of (S.33) and the third term of (S.28), we have

$$R_2 \leq \sup_{w\in\mathcal{W}}\left|\frac{AP(w)}{KL^*(w)}\right|\sup_{w\in\mathcal{W}}\left|\frac{KL^*(w)}{R(w)}\right| \leq \sup_{w\in\mathcal{W}}\left|\frac{AP(w)}{KL^*(w)}\right|\left\{1-\sup_{w\in\mathcal{W}}\left|\frac{R(w)-KL^*(w)}{KL^*(w)}\right|\right\}^{-1} \xrightarrow{P} 0.$$

For the third term R_3, we deduce that $R_3 = \left|\frac{AL(w^A)-AP(w^A)-KL(w^A)+v_n}{KL(\tilde{w}^s)}\right| \times \left|\frac{KL(\tilde{w}^s)}{KL^*(\tilde{w}^s)}\right| \times \left|\frac{KL^*(\tilde{w}^s)}{R(\tilde{w}^s)}\right| \triangleq R_{31} \times R_{32} \times R_{33}$. According to the proof of (23), we can derive that $R_{31} \xrightarrow{P} 0$. The details are omitted here. Based on the third term of (S.33), we have $R_{32} \leq 1$ with probability tending to one. Further, by the proof of $R_2 \xrightarrow{P} 0$, we can obtain $R_{33} \leq 1$ with probability tending to one. Thus, $R_3 \xrightarrow{P} 0$.

Let $o_L = \sup_{w\in\mathcal{W}}|KL(w)/KL^*(w)-1|$ and $o_R = \sup_{w\in\mathcal{W}}|R(w)/KL^*(w)-1|$. For the sufficiently large n, it is readily shown that $KL^*(w)(1-o_L) \leq KL(w) \leq KL^*(w)(1+o_L)$ and $KL^*(w)(1-o_R) \leq R(w) \leq KL^*(w)(1+o_R)$. Thus, we have $|\inf_{w\in\mathcal{W}}KL(w)-\inf_{w\in\mathcal{W}}R(w)| = o_p(\inf_{w\in\mathcal{W}}KL^*(w)) = o_p(\xi_n)$ based on the third term of (S.28) and the third term of (S.33) (which imply that $o_L \xrightarrow{P} 0$ and $o_R \to 0$). Then, for R_4, we have $R_4 \leq \frac{|\inf_{w\in\mathcal{W}}KL(w)-\inf_{w\in\mathcal{W}}R(w)|}{\inf_{w\in\mathcal{W}}KL^*(w)} \times \left|\frac{KL^*(\tilde{w}^s)}{R(\tilde{w}^s)}\right| \xrightarrow{P} 0$. Based on the above arguments, we can derive (24). This completes the proof of Theorem 4. □

References

Ando T, Li KC (2017) A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics* 45:2654–2679.

Chen Z, Liao J, Xu W, Yang Y (2023) Multifold cross-validation model averaging for generalized additive partial linear models. *Journal of Computational and Graphical Statistics* 32:1649–1659.

Cheng TC, Ing CK, Yu SH (2015) Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics* 189:321–334.

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

Li J, Lv J, Wan ATK, Liao J (2022) Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association* 117:495–509.

Liang H, Du P (2012) Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics* 6:1838–1846.

Lv J, Liu J (2014) Model selection principles in misspecified models. *Journal of the Royal Statistical Society,* Series B 76:141–167.

Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 10:1040–1053.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Whittle P (1960) Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and Its Applications* 5:302–305.

Yang W, Yang Y (2017) Toward an objective and reproducible model choice via variable selection deviation. *Biometrics* 73:20–30.

Yeh IC, Lien Ch (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36:2473–2480.

Yu D, Zhang X, Liang H (2025) Unified optimal model averaging with a general loss function based on cross-validation. *Journal of the American Statistical Association, to appear* .

Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38:894–942.

Zhang J, Chen Z, Yang Y, Xu W (2024) Variable importance based interaction modeling with an application on initial spread of COVID-19 in China. *Journal of the Royal Statistical Society, Series C* 73:1134–1154.

Zhang X, Liu CA (2023) Model averaging prediction by K-fold cross-validation. *Journal of Econometrics* 235:280–301.

Zhang X, Yu D, Zou G, Liang H (2016) Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* 111:1175–1790.

Zhang X, Zou G, Liang H, Carroll RJ (2020) Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* 115:972–984.

Zou J, Wang W, Zhang X, Zou G (2022) Optimal model averaging for divergent-dimensional Poisson regressions. *Econometric Reviews* 41:775–805.

Zou J, Wang W, Zhang X, Zou G (2025) Optimal model averaging for single index models with divergent dimensions. *Statistica Sinica* 35:1025–1049.