Supplement to “Valid Post-selection Inference in Assumption-lean Linear Regression”

Arun Kumar Kuchibhotla, Lawrence D. Brown, Andreas Buja,
Richard Berk, Linda Zhao and Edward George

31 March 2017

In this following sections, we will present complete proofs of the results stated in Sections 3 and 5 of the main paper. For the convenience in readability, we repeat the statements of the results. The supplementary material is organized as follows. We will prove Theorem 4 (of the main paper) in Section S.1 along with some generalizations. Lemmas 3 – 7 (of the main paper) in Section S.2. Theorem 5 of the main paper is proved in Section S.3.

S.1 Lasso-type Post-Selection Inference

Before proving Theorem 4 (of the main paper), we prove a simple lemma which shows that the lasso objective function is almost a surrogate loss function. Recall, we have observations $(X_i, Y_i), 1 \le i \le n$. The empirical and the expected objective functions are defined as

$$\hat{R}_M(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - X_i^\top(M)\theta\right\}^2 \quad \text{and} \quad R_M(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\{Y_i - X_i^\top(M)\theta\right\}^2\right],$$

for all $\theta \in \mathbb{R}^{|M|}$. We also have the least squares estimator and the corresponding target as

$$\beta_{0,M} := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{|M|}} R_M(\theta) \quad \text{and} \quad \hat{\beta}_M := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{|M|}} \hat{R}_M(\theta).$$

Lemma S.1.1. *For any $M \in \mathcal{M}(p)$ and $\theta \in \mathbb{R}^{|M|}$, the following inequalities hold true:*

$$\begin{aligned}\hat{R}_M(\theta) &\le R_M(\theta) + \mathcal{D}_{0n} + 2\mathcal{D}_{1n}\left\|\theta\right\|_1 + \mathcal{D}_{2n}\left\|\theta\right\|_1^2,\\ R_M(\theta) &\le \hat{R}_M(\theta) + \mathcal{D}_{0n} + 2\mathcal{D}_{1n}\left\|\theta\right\|_1 + \mathcal{D}_{2n}\left\|\theta\right\|_1^2,\end{aligned} \tag{1}$$

where

$$\mathcal{D}_{0n} := \left|\frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right]\right|.$$

Proof. We will only prove the first inequality and the second will be obvious to proof. Recall the notation,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top, \quad \text{and} \quad \Gamma_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

Similarly, $\Sigma = \mathbb{E}\left[\Sigma_n\right]$ and $\Gamma = \mathbb{E}\left[\Gamma_n\right]$. Expanding the square function in $\hat{R}_M(\theta)$, we have

$$\begin{aligned}
\hat{R}_M(\theta) &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 2\theta^\top \Gamma_n(M) + \theta^\top \Sigma_n(M)\theta \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_i^2\right] - 2\theta^\top \Gamma(M) + \theta^\top \Sigma(M)\theta + \left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right]\right) \\
&\quad + 2\theta^\top \left[\Gamma(M) - \Gamma_n(M)\right] + 2\theta^\top \left[\Sigma(M) - \Sigma_n(M)\right]\theta \\
&\le R_M(\theta) + \mathcal{D}_{0n} + 2\left\|\theta\right\|_1 \mathcal{D}_{1n} + \left\|\theta\right\|_1^2 \mathcal{D}_{2n}.
\end{aligned}$$

Combining these inequalities, we get

$$\hat{R}_M(\theta) \le R_M(\theta) + \mathcal{D}_{0n} + 2\left\|\theta\right\|_1 \mathcal{D}_{1n} + \left\|\theta\right\|_1^2 \mathcal{D}_{2n}. \qquad \square$$

Remark S.1.1 Note from the inequality (1) that the right hand side resembles the lasso objective function except for the term based on $\|\theta\|_1^2$, the coefficient of which is

$$\mathcal{D}_{2n} := \left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i X_i^\top\right]\right\|_\infty .$$

Suppose, for instance, assume that the covariates are all fixed deterministic values. Then $\mathcal{D}_{2n} = 0$ and the right hand side in inequality (1) is exactly the lasso objective function except for a constant D_{0n} that does not matter for minimizers. Also note that under setting 1(a) of Lemma 2 in the main paper, $\mathcal{D}_{1n}$ coincides in rate with the optimal tuning parameter rate in lasso. This inequality suggests that lasso is a natural candidate for high-dimensional linear regression but the derivation of a similar inequality for generalized linear models is not so obvious which questions the use of $\|\cdot\|_1$-penalty for other linear models. $\diamond$

Getting back to the lasso-type post-selection inference, for every $M \in \mathcal{M}(p)$, we have the confidence regions

$$\begin{aligned}
\check{\mathcal{R}}_M &:= \left\{\theta \in \mathbb{R}^{|M|} :\, \hat{R}_M(\theta) \le \hat{R}_M(\hat{\beta}_M) + 4C_1(\alpha)\left\|\hat{\beta}_M\right\|_1 + 2C_2(\alpha)\left\|\hat{\beta}_M\right\|_1^2\right\}, \\
\check{\mathcal{R}}_M^\dagger &:= \left\{\theta \in \mathbb{R}^{|M|} :\, \hat{R}_M(\theta) \le \hat{R}_M(\hat{\beta}_M) + 2C_1(\alpha)\left[\left\|\hat{\beta}_M\right\|_1 + \|\theta\|_1\right] + C_2(\alpha)\left[\left\|\hat{\beta}_M\right\|_1^2 + \|\theta\|_1^2\right]\right\}
\end{aligned}$$

where $\hat{R}_M(\cdot)$ is the empirical least squares objective function defined above. The following is a generalization of Theorem 4 of the main paper.

Theorem S.1.1. *Under assumptions (A2) – (A3) and for every $1 \le k \le p$ that satisfies (A4)(k), we have*

$$\liminf_{n\to\infty} \mathbb{P}\left(\bigcap_{M\in\mathcal{M}(k)} \left\{\beta_{0,M} \in \check{\mathcal{R}}_M\right\}\right) \ge 1-\alpha.$$

Under no assumptions except for (A2), we have

$$\mathbb{P}\left(\bigcap_{M\in\mathcal{M}(p)} \left\{\beta_{0,M} \in \check{\mathcal{R}}_M^{\dagger}\right\}\right) \ge 1-\alpha.$$

Proof. Since we are only concerned with objective functions at $\beta_{0,M}$, we can get a sharper inequality than the one presented in Lemma S.1.1 but the proof is similar. As noted in the main paper, we have the equalities,

$$\begin{aligned}\hat{\beta}_M &:= \arg\min_{\theta\in\mathbb{R}^{|M|}} \left\{\theta^\top \Sigma_n(M)\theta - 2\theta^\top \Gamma_n(M)\right\},\\ \beta_{0,M} &:= \arg\min_{\theta\in\mathbb{R}^{|M|}} \left\{\theta^\top \Sigma(M)\theta - 2\theta^\top \Gamma(M)\right\}.\end{aligned}$$

From these equalities, we get the following inequalities for every $M \in \mathcal{M}(p)$,

$$\begin{aligned}&\beta_{0,M}\Sigma_n(M)\beta_{0,M} - 2\beta_{0,M}^\top\Gamma_n(M)\\ &\quad\le \beta_{0,M}\Sigma(M)\beta_{0,M} - 2\beta_{0,M}^\top\Gamma(M) + 2\mathcal{D}_{1n}\left\|\beta_{0,M}\right\|_1 + \mathcal{D}_{2n}\left\|\beta_{0,M}\right\|_1^2\\ &\quad\le \hat{\beta}_M\Sigma(M)\hat{\beta}_M - 2\hat{\beta}_M^\top\Gamma(M) + 2\mathcal{D}_{1n}\left\|\beta_{0,M}\right\|_1 + \mathcal{D}_{2n}\left\|\beta_{0,M}\right\|_1^2\\ &\quad\le \hat{\beta}_M\Sigma_n(M)\hat{\beta}_M - 2\hat{\beta}_M^\top\Gamma_n(M) + 2\mathcal{D}_{1n}\left[\left\|\hat{\beta}_M\right\|_1 + \left\|\beta_{0,M}\right\|_1\right]\\ &\qquad + \mathcal{D}_{2n}\left[\left\|\hat{\beta}_M\right\|_1^2 + \left\|\beta_{0,M}\right\|_1^2\right].\end{aligned}$$

The first and third inequalities follow from the proof of Lemma S.1.1. The second inequality follows from the definition of $\beta_{0,M}$. Now add the sample average of $\{Y_i^2 : 1 \le i \le n\}$ on both sides and get

$$\hat{R}_M\left(\beta_{0,M}\right) \le \hat{R}_M\left(\hat{\beta}_M\right) + 2\mathcal{D}_{1n}\left[\left\|\hat{\beta}_M\right\|_1 + \left\|\beta_{0,M}\right\|_1\right] + \mathcal{D}_{2n}\left[\left\|\hat{\beta}_M\right\|_1^2 + \left\|\beta_{0,M}\right\|_1^2\right], \quad (2)$$

for all $M \in \mathcal{M}(p)$. Using assumptions (A4)(k) and (A3) (following the proof of Theorem 1 of the main paper), we can make the probability statement

$$\liminf_{n\to\infty} \mathbb{P}\left(\bigcap_{M\in\mathcal{M}(k)} \left\{\beta_{0,M} \in \check{\mathcal{R}}_M\right\}\right) \ge 1-\alpha.$$

The second result follows trivially from inequality (2). □

Remark S.1.2 Based on this proof, one can derive a generalization similar to Theorem 3 of the main paper with arbitrary matrices Σ_n^*, Σ^* and arbitrary vectors Γ_n^*, Γ^*. We leave it to the reader to figure out the details. $\diamond$

Interestingly, one may ask if Dantzig selector and the lasso are the only methods that can lead to valid post-selection confidence regions or is there a collection of such methods that work for this purpose. We do not yet know of any general result in this direction but we know that at least there is one another method that works. Before stating the theorem, we state a lemma that shows that square-root lasso is also a surrogate and this leads to the confidence regions.

Lemma S.1.2. *For any model $M \in \mathcal{M}(p)$ and $\theta \in \mathbb{R}^{|M|}$, the following inequality hold true:*

$$\hat{R}_M^{1/2}(\theta) \le R_M^{1/2}(\theta) + \eta_n^{1/2}\left(1 + \|\theta\|_1\right),$$
$$R_M^{1/2}(\theta) \le \hat{R}_M^{1/2}(\theta) + \eta_n^{1/2}\left(1 + \|\theta\|_1\right),$$

where $\eta_n = \max\{\mathcal{D}_{0n}, \mathcal{D}_{1n}, \mathcal{D}_{2n}\}$.

Proof. From the proof of Lemma S.1.1, we have

$$\begin{aligned}
\hat{R}_M(\theta) &\le R_M(\theta) + \mathcal{D}_{0n} + 2\mathcal{D}_{1n}\|\theta\|_1 + \mathcal{D}_{2n}\|\theta\|_1^2 \\
&\le R_M(\theta) + \eta_n\left(1 + 2\|\theta\|_1 + \|\theta\|_1^2\right) \\
&\le R_M(\theta) + \eta_n\left(1 + \|\theta\|_1\right)^2.
\end{aligned}$$

Taking square root on both sides and using the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we get

$$\hat{R}_M^{1/2}(\theta) \le R_M^{1/2}(\theta) + \eta_n^{1/2}\left(1 + \|\theta\|_1\right).$$

□

Now, consider the confidence regions

$$\breve{\mathcal{R}}_M := \left\{\theta \in \mathbb{R}^{|M|} :\ \hat{R}_M^{1/2}(\theta) \le \hat{R}_M^{1/2}(\hat{\beta}) + 2C^{1/2}(\alpha)\left(1 + \left\|\hat{\beta}_M\right\|_1\right)\right\},$$
$$\breve{\mathcal{R}}_M^\dagger := \left\{\theta \in \mathbb{R}^{|M|} :\ \hat{R}_M^{1/2}(\theta) \le \hat{R}_M^{1/2}(\hat{\beta}) + C^{1/2}(\alpha)\left(1 + \|\theta\|_1\right) + C^{1/2}(\alpha)\left(1 + \left\|\hat{\beta}_M\right\|_1\right)\right\},$$

where $C(\alpha)$ is either given by $C(\alpha) = \max\{C_1(\alpha), C_2(\alpha)\}$ or $C(\alpha)$ is the $(1-\alpha)$-upper quantile of $\max\{\mathcal{D}_{1n}, \mathcal{D}_{2n}\}$. The following theorem is the analogue of Theorem S.1.1 with square-root lasso.

Theorem S.1.2. *Under assumptions (A2) – (A3) and for every $1 \le k \le p$ that satisfies (A4)(k), we have*

$$\liminf_{n\to\infty} \mathbb{P}\left(\bigcap_{M\in\mathcal{M}(k)} \left\{\beta_{0,M} \in \breve{\mathcal{R}}_M\right\}\right) \ge 1 - \alpha.$$

Under no assumptions except for (A2), we have

$$\mathbb{P}\left(\bigcap_{M\in\mathcal{M}(p)} \left\{\beta_{0,M} \in \breve{\mathcal{R}}_M^\dagger\right\}\right) \ge 1 - \alpha.$$

Proof. The proof is almost the same as that of Theorem S.1.1 with same changes as in Lemma S.1.2. □

S.2 Uniform Convergence Results for Linear Regression

Lemma S.2.1. *If k satisfies $k\mathcal{D}_{2n} = o_p(1)$, then*

$$\sup_{M\in\mathcal{M}(k)} \|\Sigma_n(M) - \Sigma(M)\|_{op} \le k\mathcal{D}_{2n} = o_p(1).$$

Proof. Since $\Sigma_n(M) - \Sigma(M)$ is a symmetric matrix, the operator norm can be written as

$$\|\Sigma_n(M) - \Sigma(M)\|_{op} = \sup_{\delta\in\mathbb{R}^{|M|},\, \|\delta\|_2\le 1} |\delta^\top(\Sigma_n(M) - \Sigma(M))\delta|.$$

This implies that

$$\begin{aligned}
\sup_{M\in\mathcal{M}(k)} \|\Sigma_n(M) - \Sigma(M)\|_{op} &= \sup_{\substack{\delta\in\mathbb{R}^p,\\ \|\delta\|_0\le k, \|\delta\|_2\le 1}} |\delta^\top(\Sigma_n - \Sigma)\delta|,\\
&\le \sup_{\substack{\delta\in\mathbb{R}^p,\\ \|\delta\|_0\le k, \|\delta\|_2\le 1}} \|\Sigma_n - \Sigma\|_\infty \|\delta\|_1^2,\\
&\le \sup_{\substack{\delta\in\mathbb{R}^p,\\ \|\delta\|_0\le k, \|\delta\|_2\le 1}} k\|\Sigma_n - \Sigma\|_\infty \|\delta\|_2^2,\\
&= k\|\Sigma_n - \Sigma\|_\infty = k\mathcal{D}_{2n} = o_p(1),
\end{aligned}$$

by the given hypothesis. □

Remark S.2.1 To comment on how to improve this result for the case of independent and identically distributed random vectors case, we use the first equality,

$$\begin{aligned}
\sup_{M\in\mathcal{M}(k)} \|\Sigma_n(M) - \Sigma(M)\|_{op} &= \sup_{\substack{\delta\in\mathbb{R}^p,\\ \|\delta\|_0\le k, \|\delta\|_2\le 1}} |\delta^\top(\Sigma_n - \Sigma)\delta|\\
&= \sup_{\substack{\delta\in\mathbb{R}^p,\\ \|\delta\|_0\le k, \|\delta\|_2\le 1}} \left|\frac{1}{n}\sum_{i=1}^n \delta^\top(X_iX_i^\top - \Sigma)\delta\right|.
\end{aligned}$$

Now, use the techniques of symmetrization using Rademacher variables and covering numbers for sparse vectors to get the correct rate. See Rudelson and Vershynin (2008) for more details. ◇

Before proceeding to prove the remaining uniform consistency results, we need a result that proves closeness of minimizers of close convex functions which was proved by Hjort and Pollard (1993). For completeness, we state the result along with a detailed proof here. As will be seen from the proof of the lemma, there is no randomness involved.

Lemma S.2.2. *[Lemma 2 of Hjort and Pollard (1993)] Suppose $\{A_n(s)\}$ is a sequence of convex functions defined on an open convex set $\mathcal{S}$ and $\{B_n(s)\}$ is another sequence of functions. Let α_n be the minimizer of A_n and assume that B_n has a unique global minimizer β_n. Then for each $\delta > 0$ and any norm $|\cdot|$,*

$$\mathbb{P}\left(|\alpha_n - \beta_n| \geq \delta\right) \leq \mathbb{P}\left(\Delta_n(\delta) \geq \frac{1}{2}h_n(\delta)\right),$$

where

$$\Delta_n(\delta) := \sup_{|s-\beta_n|\leq\delta} |A_n(s) - B_n(s)| \quad \textit{and} \quad h_n(\delta) := \inf_{|s-\beta_n|=\delta} B_n(s) - B_n(\beta_n).$$

Proof. Define $r_n(s) := A_n(s) - B_n(s)$ and we have the minimizers

$$\alpha_n = \arg\min_{\theta\in\mathcal{S}} A_n(\theta) \quad \text{and} \quad \beta_n = \arg\min_{\theta\in\mathcal{S}} B_n(\theta).$$

Let s be any point in $\mathcal{S}$ outside the ball around β_n with radius δ, say, $s = \beta_n + \ell u$ with $|u| = 1$ and $\ell > \delta$. By convexity of A_n, we get

$$A_n(\beta_n + \delta u) = A_n\left(\left[1 - \frac{\delta}{\ell}\right]\beta_n + \frac{\delta}{\ell}s\right) \leq \left(1 - \frac{\delta}{\ell}\right)A_n(\beta_n) + \frac{\delta}{\ell}A_n(s).$$

Rearranging, we arrive at the inequality

$$\begin{aligned}\frac{\delta}{\ell}\left\{A_n(s) - A_n(\beta_n)\right\} &\geq A_n(\beta_n + \delta u) - A_n(\beta_n)\\ &= B_n(\beta_n + \delta u) - B_n(\beta_n) + r_n(\beta_n + \delta u) - r_n(\beta_n)\\ &\geq \inf_{|u|=1} B_n(\beta_n + \delta u) - B_n(\beta_n) - 2\sup_{|v|\leq 1}|r_n(\beta_n + \delta v)|\\ &\geq h_n(\delta) - 2\Delta_n(\delta).\end{aligned}$$

If $h_n(\delta) \geq 2\Delta_n(\delta)$, then $A_n(s) \geq A_n(\beta_n)$ for all s satisfying $|s - \beta_n| \geq \delta$ and so the minimizer of A_n cannot lie outside the ball around β_n of radius δ. Therefore,

$$\{|\alpha_n - \beta_n| \geq \delta\} \subseteq \{h_n(\delta) \leq 2\Delta_n(\delta)\},$$

and implies the result. □

It is easy to note from the inequality that if A_n and B_n are uniformly close and $1/h_n(\delta)$ is bounded, then α_n and β_n are also close. It would also be convenient (notationally) to define errors for model M, even though these errors are not assumed to have any special properties: For any model M,

$$\varepsilon_{i,M} := Y_i - X_i^\top(M)\beta_{0,M} \quad \text{and} \quad \varepsilon_M := Y - X^\top(M)\beta_{0,M}.$$

Here again we are writing M in the subscript to re-emphasize the fact that $\varepsilon_{i,M}$ is not an element of a fixed big vector of length p.

Lemma S.2.3. *For any $k \geq 1$ and $n \geq 1$ that satisfy $2k\mathcal{D}_{2n} \leq \lambda_{\min}(\Sigma)$, the following stochastic ordering holds:*

$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_M - \beta_{0,M} \right\|_1 \preceq \frac{4k(\mathcal{D}_{1n} + \mathcal{D}_{2n} S_{1,k})}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}} = o_p(1),$$

where the last equality holds under assumptions (A3) and (A4)(k) in main paper.

Remark S.2.2 We provide two different proofs of Lemma S.2.3. The first proof is very specific to the case of linear regression and the second proof even though long generalizes to the other M-estimation problems easily. See Negahban et al. (2009) for more details. ◇

Proof 1 of Lemma S.2.3. The least squares estimator $\hat{\beta}_M$ satisfies

$$\Sigma_n(M)\hat{\beta}_M - \Gamma_n(M) = 0,$$

and so, we have

$$\hat{\beta}_M - \beta_{0,M} = (\Sigma_n(M))^{-1} \left([\Gamma_n(M) - \Gamma(M)] - [\Sigma_n(M) - \Sigma(M)]\, \beta_{0,M} \right).$$

By Lemma S.2.1, we have

$$\|\Sigma(M)\|_{op} - k\mathcal{D}_{2n} \leq \|\Sigma_n(M)\|_{op} \leq \|\Sigma(M)\| + k\mathcal{D}_{2n}.$$

Therefore, for k satisfying $k\mathcal{D}_{2n} \leq \lambda_{\min}(\Sigma)$,

$$\left\| \hat{\beta}_M - \beta_{0,M} \right\|_2 \leq \frac{\|\Gamma_n(M) - \Gamma(M)\|_2 + \|[\Sigma_n(M) - \Sigma(M)]\beta_{0,M}\|_2}{\lambda_{\min}(\Sigma(M)) - k\mathcal{D}_{2n}}.$$

Using $\|\cdot\|_2 - \|\cdot\|_1$ inequality, we obtain uniformly over $M \in \mathcal{M}(k)$,

$$\left\| \hat{\beta}_M - \beta_{0,M} \right\|_1 \leq k^{1/2} \left\| \hat{\beta}_M - \beta_{0,M} \right\|_2 \leq \frac{k(\mathcal{D}_{1n} + \mathcal{D}_{2n} S_{1,k})}{\lambda_{\min}(\Sigma(M)) - k\mathcal{D}_{2n}}.$$

Note that this improves the results with respect to constants. □

Proof 2 of Lemma S.2.3. A naive application of Lemma S.2.2 with $A_n(\cdot)$ and $B_n(\cdot)$ replaced by $\hat{R}_M(\cdot)$ and $R_M(\cdot)$ will bring in the estimation error of the sample mean of Y_i^2 into the conditions for uniform consistency required. However, this term can be avoided by realizing that the least squares estimator $\hat{\beta}_M$ is also the minimizer of the least squares loss $\mathcal{O}_n(s; M)$ subtracted by the mean of Y_i^2, $1 \leq i \leq n$ as stated after Lemma 1 in the main paper.

Note that for $M \in \mathcal{M}(k)$, $\hat{\gamma}_M := \left(\hat{\beta}_M - \beta_{0,M} \right)$ is the minimizer of

$$\begin{aligned} \hat{R}_M(\beta_{0,M} + s) - \hat{R}_M(\beta_{0,M}) = {} & -2\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top(M)\beta_{0,M})X_i^\top(M)s \\ & + s^\top \left(\frac{1}{n}\sum_{i=1}^{n} X_i(M)X_i^\top(M) \right) s. \end{aligned}$$

For $M \in \mathcal{M}(k)$ and $s \in \mathbb{R}^{|M|}$, define the objective functions

$$A_n(s; M) := \hat{R}_M(\beta_{0,M} + s) - \hat{R}_M(\beta_{0,M}) = -2\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,M} X_i^\top(M)s + s^\top \Sigma_n(M)s.$$

$$B_n(s; M) := -2\mathbb{E}\left[\varepsilon_M X^\top(M)\right] s + s^\top \Sigma(M)s = s^\top \Sigma(M)s.$$

The second equality in the definition of $B_n(s; M)$ holds by the definition of $\beta_{0,M}$. It is easy to see that the minimizer of $B_n(s; M)$ with respect to $s \in \mathbb{R}^{|M|}$ is $0 \in \mathbb{R}^{|M|}$.

For applying Lemma S.2.2, define for $M \in \mathcal{M}(k)$,

$$\Delta_n(\delta; M) = \sup_{s \in \mathbb{R}^{|M|}:\, \|s\|_1 \le \delta} |A_n(s; M) - B_n(s; M)|,$$

$$h_n(\delta; M) = \inf_{s \in \mathbb{R}^{|M|}:\, \|s\|_1 = \delta} B_n(s).$$

The uniform consistency of $\hat{\beta}_M$ to $\beta_{0,M}$ over $M \in \mathcal{M}(k)$ is equivalent to proving that for any $\delta > 0$,

$$\mathbb{P}\left(\sup_{M \in \mathcal{M}(k)} \|\hat{\gamma}_M\|_1 \ge \delta\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Notice that from the proof of Lemma S.2.2,

$$\begin{aligned}\left\{\sup_{M \in \mathcal{M}(k)} \|\hat{\gamma}_M\|_1 \ge \delta\right\} &= \left\{\sup_{M \in \mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \ge \delta\right\} \\ &= \bigcup_{M \in \mathcal{M}(k)} \left\{\left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \ge \delta\right\} \\ &\subseteq \bigcup_{M \in \mathcal{M}(k)} \left\{h_n(\delta; M) \le 2\Delta_n(\delta; M)\right\}. \end{aligned} \tag{3}$$

Firstly, we have

$$B_n(s) \ge \lambda_{\min}(\Sigma(M)) \|s\|_2^2 \ge \frac{\lambda_{\min}(\Sigma)}{|M|} \|s\|_1^2 \quad \Rightarrow \quad h_n(\delta; M) \ge \frac{\lambda_{\min}(\Sigma)}{|M|}\delta^2 \tag{4}$$

To deal with $\Delta_n(\delta; M)$, note that

$$\begin{aligned}|A_n(s; M) - B_n(s; M)| \le 2 &\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,M} X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_\infty \|s\|_1 \\ &+ \|\Sigma_n(M) - \Sigma(M)\|_\infty \|s\|_1^2 .\end{aligned}$$

This implies under the condition $\|s\|_1 \le \delta$,

$$\Delta_n(\delta; M) \le 2\left\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,M} X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_\infty \delta + \|\Sigma_n(M) - \Sigma(M)\|_\infty \delta^2. \tag{5}$$

Breaking down the $\|\cdot\|_\infty$-norm in the first term by substituting the definitions of $\varepsilon_{i,M}$ and ε_M, we see that

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_\infty\\ &\quad\le \left\|\frac{1}{n}\sum_{i=1}^n X_i(M)Y_i - \mathbb{E}\left[X(M)Y\right]\right\|_\infty + \left\|\left[\Sigma_n(M) - \Sigma(M)\right]\beta_{0,M}\right\|_\infty\\ &\quad\le \left\|\frac{1}{n}\sum_{i=1}^n X_iY_i - \mathbb{E}\left[XY\right]\right\|_\infty + \left\|\Sigma_n(M) - \Sigma(M)\right\|_\infty \left\|\beta_{0,M}\right\|_1 \le \mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}.\end{aligned}$$

Hence,

$$\Delta_n(\delta;M) \le 2(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\delta + \mathcal{D}_{2n}\delta^2. \tag{6}$$

Substituting these inequalities in the inclusion (3), we get

$$\begin{aligned}\left\{\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_1 \ge \delta\right\} &\subseteq \bigcup_{M\in\mathcal{M}(k)} \left\{h_n(\delta;M) \le 2\Delta_n(\delta;M)\right\}\\ &\subseteq \bigcup_{M\in\mathcal{M}(k)} \left\{\frac{\lambda_{\min}(\Sigma)}{|M|}\delta^2 \le 4(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\delta + 2\mathcal{D}_{2n}\delta^2\right\} &(7)\\ &\subseteq \bigcup_{M\in\mathcal{M}(k)} \left\{\frac{\lambda_{\min}(\Sigma)}{2|M|}\delta^2 \le 4(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\delta\right\} \cup\\ &\qquad \bigcup_{M\in\mathcal{M}(k)} \left\{\frac{\lambda_{\min}(\Sigma)}{2|M|}\delta^2 \le 2\mathcal{D}_{2n}\delta^2\right\}\\ &\subseteq \left\{\lambda_{\min}(\Sigma)\delta \le 8k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\right\} \cup \left\{\lambda_{\min}(\Sigma) \le 4k\mathcal{D}_{2n}\right\}. &(8)\end{aligned}$$

Therefore, by union bound we get

$$\mathbb{P}\left(\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_1 \ge \delta\right) \le \mathbb{P}\left(8k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}) \ge \lambda_{\min}(\Sigma)\delta\right) + \mathbb{P}\left(4k\mathcal{D}_{2n} \ge \lambda_{\min}(\Sigma)\right).$$

Hence the uniform rate of consistency holds, i.e.,

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 = O_p\left(k\mathcal{D}_{1n} + k\mathcal{D}_{2n}S_{1,k}\right), \quad \text{if} \quad k\mathcal{D}_{2n} = o_p(1).$$

From the inclusion (7), we have

$$\begin{aligned}\left\{\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_1 \ge \delta\right\} &\subseteq \bigcup_{M\in\mathcal{M}(k)} \left\{\frac{\lambda_{\min}(\Sigma)}{|M|}\delta^2 \le 4(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\delta + 2\mathcal{D}_{2n}\delta^2\right\}\\ &\subseteq \left\{\left(\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}\right)\delta \le 4k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})\right\}.\end{aligned}$$

Thus, we obtain for every $\delta > 0$ and every k such that $2k\mathcal{D}_{2n} < \lambda_{\min}(\Sigma)$,

$$\mathbb{P}\left(\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \ge \delta\right) \le \mathbb{P}\left(\frac{4k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}} \ge \delta\right),$$

and so,

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \preceq \frac{4k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}}.$$

□

Remark S.2.3 Improvement to the case of independent and identically distributed random vectors uses symmetrization techniques from the inequality (5). ◇

The complimentary results for uniform consistency in $\|\cdot\|_2$-norm follows. Here too two different proofs are possible. We do not repeat the proof 1 which is similar to that of Lemma S.2.3.

Lemma S.2.4. *For any $k \ge 1$, and $n \ge 1$ that satisfy $2k\mathcal{D}_{2n} \le \lambda_{\min}(\Sigma)$, the following stochastic ordering holds:*

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_2 \preceq \frac{4(k^{1/2}\mathcal{D}_{1n} + RIP_n(k)S_{2,k})}{\lambda_{\min}(\Sigma) - 2RIP_n(k)} \preceq \frac{4(k^{1/2}\mathcal{D}_{1n} + k\mathcal{D}_{2n}S_{2,k})}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}}.$$

Proof. We follow the proof of Lemma S.2.3 with $\|\cdot\|_2$-norm replacing the $\|\cdot\|_1$-norm. Let $A_n(\cdot; M)$ and $B_n(\cdot; M)$ be the same functions defined in Lemma S.2.3. For applying Lemma S.2.2, define for $M \in \mathcal{M}(k)$,

$$\begin{aligned}\Delta_n(\delta; M) &= \sup_{s\in\mathbb{R}^{|M|}:\, \|s\|_2\le\delta} |A_n(s; M) - B_n(s; M)|,\\ h_n(\delta; M) &= \inf_{s\in\mathbb{R}^{|M|}:\, \|s\|_2=\delta} B_n(s).\end{aligned}$$

As in the proof of Lemma S.2.3, we have the inclusion,

$$\begin{aligned}\left\{\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_2 \ge \delta\right\} &= \left\{\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_2 \ge \delta\right\}\\ &\subseteq \bigcup_{M\in\mathcal{M}(k)} \{h_n(\delta; M) \le 2\Delta_n(\delta; M)\}\\ &\subseteq \left\{\inf_{M\in\mathcal{M}(k)} h_n(\delta; M) \le \sup_{M\in\mathcal{M}(k)} 2\Delta_n(\delta; M)\right\}. \qquad (9)\end{aligned}$$

By definition of the minimum eigenvalue, we have

$$h_n(\delta; M) = \lambda_{\min}(\Sigma(M))\delta^2 \quad \Rightarrow \quad \inf_{M\in\mathcal{M}(k)} h_n(\delta; M) \ge \lambda_{\min}(\Sigma)\delta^2.$$

It is clear by Cauchy-Schwarz inequality that

$$\Delta_n(\delta; M) \le 2\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_2 \delta + \delta^2 \left\|\Sigma_n(M) - \Sigma(M)\right\|_{op}.$$

This implies (using (9)) that

$$\mathbb{P}\left(\sup_{M\in\mathcal{M}(k)}\|\hat{\gamma}_M\|_2 \ge \delta\right) \le \mathbb{P}\left(\inf_{M\in\mathcal{M}(k)} h_n(\delta; M) \le 2\sup_{M\in\mathcal{M}(k)}\Delta_n(\delta; M)\right).$$

Using the two-term bound on $\Delta_n(\delta; M)$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{M\in\mathcal{M}(k)}\|\hat{\gamma}_M\|_2 \ge \delta\right)\\
&\qquad\le \mathbb{P}\left(4\sup_{M\in\mathcal{M}(k)}\|\Sigma_n(M)-\Sigma(M)\|_{op} \ge \lambda_{\min}(\Sigma)\right)\\
&\qquad+\mathbb{P}\left(8\sup_{M\in\mathcal{M}(k)}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_2 \ge \lambda_{\min}(\Sigma)\delta\right)\\
&\qquad\le \mathbb{P}\left(4k\mathcal{D}_{2n} \ge \lambda_{\min}(\Sigma)\right)\\
&\qquad+\mathbb{P}\left(8\sup_{M\in\mathcal{M}(k)}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_2 \ge \lambda_{\min}(\Sigma)\delta\right),
\end{aligned}$$

where the last inequality follows from the calculations in the proof of Lemma S.2.1. This inequality implies that if $k\mathcal{D}_{2n} = o_p(1)$, then

$$\sup_{M\in\mathcal{M}(k)}\|\hat{\gamma}_M\|_2 = O_p\left(\sup_{M\in\mathcal{M}(k)}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_2\right).$$

This equality is rate-sharp in the sense that if the supremum were absent, then upto the rate the equality is exact. To be more transparent, we now simplify the supremum term on the right hand side above. We have the following sequence of inequalities,

$$\begin{aligned}
&\sup_{M\in\mathcal{M}(k)}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,M}X_i(M) - \mathbb{E}\left[\varepsilon_M X(M)\right]\right\|_2\\
&\le \sup_{M\in\mathcal{M}(k)}\left\|\frac{1}{n}\sum_{i=1}^n X_i(M)Y_i - \mathbb{E}\left[X(M)Y\right]\right\|_2 + \sup_{M\in\mathcal{M}(k)}\|\left[\Sigma_n(M)-\Sigma(M)\right]\beta_{0,M}\|_2\\
&\le k^{1/2}\left\|\frac{1}{n}\sum_{i=1}^n X_iY_i - \mathbb{E}\left[XY\right]\right\|_\infty + \sup_{M\in\mathcal{M}(k)}\|\Sigma_n(M)-\Sigma(M)\|_{op}\|\beta_{0,M}\|_2\\
&\le k^{1/2}\mathcal{D}_{1n} + \mathrm{RIP}_n(k)S_{2,k} \le k^{1/2}\mathcal{D}_{1n} + k\mathcal{D}_{2n}S_{2,k}.
\end{aligned}$$

The last inequality here follows from Lemma S.2.1. Therefore,

$$\sup_{M\in\mathcal{M}(k)}\|\hat{\gamma}_M\|_2 = O_p\left(k^{1/2}\mathcal{D}_{1n} + \mathrm{RIP}_n(k)S_{2,k}\right) \le O_p(k^{1/2}\mathcal{D}_{1n} + k\mathcal{D}_{2n}S_{2,k}),$$

where the last inequality follows from $\mathrm{RIP}_n(k) \le k\mathcal{D}_{2n}$. Recall $S_{2,k} = \sup_{M\in\mathcal{M}(k)} \|\beta_{0,M}\|_2$. This rate is in accordance with the rate obtained in Lemma S.2.3 in the sense that we got

$$\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_2 = O_p\left(k^{-1/2} \sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_1\right),$$

where we used the fact that $S_{1,k} \le k^{1/2} S_{2,k}$. Following the final steps in the proof of Lemma S.2.3, we obtain the inequality

$$\mathbb{P}\left(\sup_{M\in\mathcal{M}(k)} \|\hat{\gamma}_M\|_2 \ge \delta\right) \le \mathbb{P}\left(\frac{4(k^{1/2}\mathcal{D}_{1n} + \mathrm{RIP}_n(k)S_{2,k})}{\lambda_{\min}(\Sigma) - 2\mathrm{RIP}_n(k)} \ge \delta\right),$$

for every $\delta > 0$ and k such that $2\mathrm{RIP}_n(k) \le \delta$ and hence,

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_2 \preceq \frac{4(k^{1/2}\mathcal{D}_{1n} + \mathrm{RIP}_n(k)S_{2,k})}{\lambda_{\min}(\Sigma) - 2\mathrm{RIP}_n(k)} \preceq \frac{4(k^{1/2}\mathcal{D}_{1n} + k\mathcal{D}_{2n}S_{2,k})}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}}. \qquad \square$$

The constants 4 and 2 above can be removed by using the inequalities in proof 1 of Lemma S.2.3.

Lemma S.2.5. *For any $k \ge 1$ such that assumptions (A3) and (A4)(k) are satisfied, the uniform relative Lebesgue measure result holds:*

$$\sup_{M\in\mathcal{M}(k)} \frac{\nu(\hat{\mathcal{R}}_M)}{(C_1(\alpha) + C_2(\alpha)S_{1,k})^{|M|}} = O_p(1).$$

Hence, it can be said that $\nu(\hat{\mathcal{R}}_M) = O_p(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})^{|M|}$ uniformly for $M \in \mathcal{M}(k)$. Moreover, additionally under the setting 1(a) of Lemma 2, we have

$$\nu\left(\hat{\mathcal{R}}_M\right) = O_p\left(k\sqrt{\frac{\log p}{n}}\right)^{|M|} \qquad \text{uniformly for } M \in \mathcal{M}(k).$$

Proof. For any fixed model M, the Lebesgue measure of the confidence region is given by

$$\nu(\hat{\mathcal{R}}_M) = |\Sigma_n^{-1}(M)| \left(C_1(\alpha) + C_2(\alpha)\left\|\hat{\beta}_M\right\|_1\right)^{|M|}, \tag{10}$$

which converges to zero as n tends to infinity. Here $\nu(A)$ is used to denote the Lebesgue measure of the set A (ν can be over different dimensions) and for any matrix $A \in \mathbb{R}^{p\times p}$, $|A|$ denotes the determinant of A. This equality follows since the confidence region $CI(M)$ can be written as

$$\hat{\mathcal{R}}_M = \left\{\Sigma_n^{-1}(M)(\theta + \hat{\beta}_M) :\ \|\theta\|_\infty \le \left(C_1(\alpha) + C_2(\alpha)\left\|\hat{\beta}_M\right\|_1\right)\right\}.$$

By Lemma S.2.1, we get with probability converging to one,

$$\sup_{M\in\mathcal{M}(k)} |\Sigma_n^{-1}(M)| \le 2\left(1 + \max_{M\in\mathcal{M}(k)} |\Sigma^{-1}(M)|\right).$$

We know that $C_1(\alpha)$ and $C_2(\alpha)$ converge to zero for all distributions with finite fourth moment with exact rate depending on how thin the tails are of the whole distribution. Under the conditions of Lemma S.2.1, the sequence $C_1(\alpha)+C_2(\alpha)S_{1,k}$ converges to zero as $n \to \infty$. The result now follows from equation (10) and uniform consistency of $\hat{\beta}_M$ in the $\|\cdot\|_1$-norm.

The second result follows by plugging-in the result of Lemma 2, setting 1(a). □

Lemma S.2.6. *Under the condition $\mathcal{D}_{0n} = o_p(1)$, assumptions (A3) and (A4)(k), we have*

$$\sup_{M\in\mathcal{M}(k)} \left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op} = k^{3/2} \max_{1\le i\le n} \|X_i\|_\infty^2 \, O_p\left(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}\right) = o_p(1).$$

Furthermore, if $X_i(j)$ is sub-Gaussian for every $1 \le j \le n$ as in setting 1(a) of Lemma 2, then

$$\sup_{M\in\mathcal{M}(k)} \left\|V_n(M) - \tilde{V}_n(M)\right\|_{op} = k^{3/2}\log(pn) O_p(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}).$$

If instead we have bounded covariates, then

$$\sup_{M\in\mathcal{M}(k)} \left\|V_n(M) - \tilde{V}_n(M)\right\|_{op} = k^{3/2} O_p(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}).$$

In fact the following exact stochastic ordering is true. Suppose there exists a real constant μ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y^2\right] \le \mu^2 < \infty.$$

For every $k \ge 1$ and $n \ge 1$, that satisfy $2k\mathcal{D}_{2n} \le \lambda_{\min}(\Sigma)$, the following stochastic ordering holds true:

$$\sup_{M\in\mathcal{M}(k)} \left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op} \preceq 8k \max_{1\le i\le n} \|X_i\|_\infty^2 \, \mathcal{E}_{n,k}^{1/2} \left[\mu + \mathcal{E}_{n,k}^{1/2} + 4(1+S_{1,k}) \left\|\Omega_n - \Omega\right\|_\infty\right],$$

where

$$\mathcal{E}_{n,k} := \frac{k(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})^2}{\lambda_{\min}(\Sigma) - 2k\mathcal{D}_{2n}}.$$

Proof. We will only prove the last result since the first two follow from the last by using the fact that maximum of N sub-Gaussian (possibly dependent) random variables is of order $\sqrt{\log N}$. Also, we assume identical distribution since the generalization to non-identical distribution follows trivially.

The difference $V_{n,M} - \tilde{V}_{n,M}$ can be written as

$$V_{n,M} - \tilde{V}_{n,M} = \frac{1}{n}\sum_{i=1}^{n} X_i(M)X_i^\top(M)\left[(Y_i - X_i^\top(M)\hat{\beta}_M)^2 - (Y_i - X_i^\top(M)\beta_{0,M})^2\right].$$

Firstly note that $\left\|X_i(M)X_i^\top(M)\right\|_{op} = \|X_i(M)\|_2^2$ and by triangle inequality of operator norm, we obtain

$$\left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op} \le \max_{1\le i\le n} \|X_i(M)\|_2^2 \frac{1}{n}\sum_{i=1}^n \left|(Y_i - X_i^\top(M)\hat{\beta}_M)^2 - (Y_i - X_i^\top(M)\beta_{0,M})^2\right|.$$

For any two real numbers A, B and any $L > 0$, we have $2|AB| \le LA^2 + L^{-1}B^2$ and so,

$$|a^2 - b^2| = |(a-b)^2 + 2(a-b)b| \le (1+L)(a-b)^2 + L^{-1}b^2 \quad \text{for all} \quad a, b \in \mathbb{R}.$$

Applying this inequality with $a = Y_i - X_i^\top(M)\hat{\beta}_M$ and $b = Y_i - X_i^\top(M)\beta_{0,M}$, we get for any $L > 0$,

$$\begin{aligned}\left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op} &\le \max_{1\le i\le n} \|X_i(M)\|_2^2 \frac{(1+L)}{n}\sum_{i=1}^n \left\{X_i^\top(M)(\hat{\beta}_M - \beta_{0,M})\right\}^2 \\ &\quad + \max_{1\le i\le n} \|X_i(M)\|_2^2 \frac{1}{nL}\sum_{i=1}^n (Y_i - X_i^\top(M)\beta_{0,M})^2. \qquad (11)\end{aligned}$$

From the definition of $\hat{\beta}_{0,M}$, we get

$$\begin{aligned}&\sum_{i=1}^n (Y_i - X_i^\top(M)\hat{\beta}_M)^2 \le \sum_{i=1}^n (Y_i - X_i^\top(M)\beta_{0,M})^2 \\ \Rightarrow &\sum_{i=1}^n \left\{X_i^\top(M)(\hat{\beta}_M - \beta_{0,M})\right\}^2 \le 2\sum_{i=1}^n (Y_i - X_i^\top(M)\beta_{0,M})X_i^\top(M)(\hat{\beta}_M - \beta_{0,M}).\end{aligned}$$

Using the Cauchy-Schwarz inequality with $\|\cdot\|_1 - \|\cdot\|_\infty$-norms, we get

$$\frac{1}{n}\sum_{i=1}^n \left\{X_i^\top(M)(\hat{\beta}_M - \beta_{0,M})\right\}^2 \le 2(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}) \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1,$$

by noting that $\mathbb{E}\left[(Y - X^\top(M)\beta_{0,M})X(M)\right] = 0$ and using the inequalities in the proof of Theorem 1 of the main paper. Substituting these bounds in inequality (11), we have

$$\begin{aligned}\left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op} &\le (2+2L)\max_{1\le i\le n} \|X_i(M)\|_2^2 (\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}) \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \\ &\quad + \max_{1\le i\le n} \|X_i(M)\|_2^2 \frac{1}{nL}\sum_{i=1}^n (Y_i - X_i^\top(M)\beta_{0,M})^2.\end{aligned}$$

Now, minimizing over $L > 0$, we arrive at the inequality

$$\begin{aligned}\frac{\left\|V_{n,M} - \tilde{V}_{n,M}\right\|_{op}}{\max_{1\le i\le n}\|X_i(M)\|_2^2} &\le 2(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k}) \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1 \\ &\quad + 2\sqrt{2}(\mathcal{D}_{1n} + \mathcal{D}_{2n}S_{1,k})^{1/2} \left\|\hat{\beta}_M - \beta_{0,M}\right\|_1^{1/2} \hat{R}_M(\beta_{0,M}). \qquad (12)\end{aligned}$$

To deal with the last factor on the second term above, observe that

$$\begin{aligned}
\hat{R}_M^{1/2}(\beta_{0,M}) &\le \left|\hat{R}_M(\beta_{0,M}) - R_M(\beta_{0,M})\right|^{1/2} + R_M^{1/2}(\beta_{0,M}) \\
&\le \left|\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}\left[Y^2\right]\right|^{1/2} + \sqrt{2}(\mathcal{D}_{1n} + \mathcal{D}_{2n}\left\|\beta_{0,M}\right\|_1)^{1/2}\left\|\beta_{0,M}\right\|_1^{1/2} \\
&\qquad + R_M^{1/2}(\beta_{0,M}) \\
&\le \mathcal{D}_{0n}^{1/2} + \sqrt{2}(\mathcal{D}_{1n} + \mathcal{D}_{1n}S_{1,k})^{1/2}S_{1,k}^{1/2} + R_M^{1/2}(\beta_{0,M}),
\end{aligned}$$

where the second inequality follows by definition of $\mathcal{D}_{1n}$ and $\mathcal{D}_{2n}$. Also, note that the first two terms on the right hand side converge to zero as $n \to \infty$. By the definition of $\beta_{0,M}$, we have

$$R_M^{1/2}(\beta_{0,M}) = \min_{s\in\mathbb{R}^{|M|}} \mathbb{E}\left[(Y - X^\top(M)s)^2\right] \le \mathbb{E}\left[Y^2\right] \Rightarrow \sup_{M\in\mathcal{M}(k)} \mathcal{O}(\beta_{0,M}; M) \le \mathbb{E}\left[Y^2\right].$$

We, of course, need the assumption of finite second moment Y to consider best linear regression functional. Substituting this bound in (12) and using the result of Lemma S.2.3, the required result is proved by noting that $\max\{\mathcal{D}_{0n}, \mathcal{D}_{1n}, \mathcal{D}_{2n}\} = \left\|\Omega_n - \Omega\right\|_\infty$ and

$$\max_{1\le i\le n} \left\|X_i(M)\right\|_2^2 \le k \max_{1\le i\le n} \left\|X_i\right\|_\infty^2 = k \max_{1\le i\le n; 1\le j\le p} |X_i(j)|^2. \qquad \square$$

S.3 Linear Regression under IID Setting

The following proof is a bit notationally involved as we need to prove semi-parametric efficiency. The first two parts of the theorem can be generalized to the case of non-iid random vectors as long as we can prove a weak law of large number and a central limit theorem.

Theorem S.3.1. *Suppose that the observations (X_i, Y_i) are independent and identically distributed with finite fourth moments. For any fixed model M of size not changing with n, under assumption (A3), we have the following results*

(a) *The least squares estimator $\hat{\beta}_M$ converges in probability to $\beta_{0,M}$ as $n \to \infty$;*

(b) *The following asymptotic normality result holds:*

$$n^{1/2}\left(\hat{\beta}_M - \beta_{0,M}\right) \xrightarrow{\mathcal{L}} N_{|M|}\left(0, B_M^{-1}V_M B_M^{-1}\right),$$

where $B_M = \Sigma(M)$ and $V_M = \mathbb{E}\left[X(M)X^\top(M)\{Y - X^\top(M)\beta_{0,M}\}^2\right]$.

(c) *The estimator $\hat{\beta}_M$ is a semi-parametrically efficient estimator of $\beta_{0,M}$ over all distributions of (X_1, Y_1) with finite fourth moments.*

(d) *The well-known sandwich estimator* $[\Sigma_n(M)]^{-1} V_{n,M} [\Sigma_n(M)]^{-1}$ *is a consistent estimator of* $B_M^{-1} V_M B_M^{-1}$*, where*

$$V_{n,M} = \frac{1}{n}\sum_{i=1}^{n} X_i(M) X_i^\top(M)\{Y_i - X_i^\top(M)\hat{\beta}_M\}^2.$$

Moreover, $[\Sigma_n(M)]^{-1} V_{n,M} [\Sigma_n(M)]^{-1}$ *is a semi-parametrically efficient estimator of* $B_M^{-1} V_M B_M^{-1}$.

Proof. Since the result is about least squares estimator for a fixed M of fixed cardinality (not changing with n), let $W_i = X_i(M)$ for $1 \le i \le n$. Let the regression estimators for these variables be given by

$$\hat{\alpha} := \arg\min_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - W_i^\top \theta\right)^2 \quad \text{and} \quad \alpha := \arg\min_{\theta \in \mathbb{R}^{|M|}} \mathbb{E}\left[(Y - W^\top\theta)^2\right].$$

By definition, we get that

$$\frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \hat{\alpha}) = 0.$$

Now adding and subtracting α to $\hat{\alpha}$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \alpha) = \frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top (\hat{\alpha} - \alpha).$$

Under finite fourth moments of the variables, $W_i(Y_i - W_i^\top \alpha)$ for $1 \le i \le n$ have finite second moments. Hence, by the weal law of large numbers, we get that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \alpha) \xrightarrow{P} 0.$$

The mean of the right hand side above is zero by definition of α. Again by weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top \xrightarrow{P} \mathbb{E}\left[WW^\top\right].$$

Since the matrices here are finite dimensional and all norms are equivalent, we have positive definiteness of the sample mean of $W_i W_i^\top$ for large enough n with probability converging to one and thus,

$$\left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top\right)^{-1} \xrightarrow{P} \left(\mathbb{E}\left[WW^\top\right]\right)^{-1}.$$

Combining these probability convergences, $\hat{\alpha}$ converges in probability to α as $n \to \infty$.

By the classical central limit theorem, we also can derive

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \alpha) \xrightarrow{\mathcal{L}} N_{|M|}\left(0, V_M\right).$$

Since

$$n^{1/2}\left(\hat{\alpha}-\alpha\right) = \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \alpha),$$

by an application of Slutsky theorem, we can write the asymptotic linear representation

$$n^{1/2}\left(\hat{\alpha}-\alpha\right) = \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(Y_i - W_i^\top \alpha) + o_p(1),$$

which then proves the asymptotic normality of $n^{1/2}(\hat{\alpha}-\alpha)$.

To prove semi-parametric efficiency of the estimator $\hat{\alpha}$: Suppose the true distribution of $(W,Y) \in \mathbb{R}^{|M|}\times\mathbb{R}$ is P and let $(w,y)\mapsto g(w,y)$ be any map such that $\mathbb{E}_P g = 0$ and $\mathbb{E}_P g^2 < \infty$. Define the one-dimensional parametric family indexed by a variable t as

$$dP_t(w,y) = c(t)K(tg(w,y))dP(w,y), \quad \text{with} \quad K(u) = 2(1+\exp(-2u))^{-1},$$

and $c(t)$ such that P_t defines a probability measure. The family $\{P_t : t \in \mathbb{R}\}$ is quadratic mean differentiable and the map g is the score function corresponding to at $t = 0$. Any function in the set $\{(x,y)\mapsto g(w,y) : \mathbb{E}_P g = 0, \mathbb{E}_P g^2 < \infty\}$ is a valid score function and defines a valid probability family. We want to estimate the least squares slope functional given by

$$\alpha = \psi(P) = \left(\mathbb{E}\left[WW^\top\right]\right)^{-1}\mathbb{E}\left[WY\right] = \left(\int ww^\top dP\right)^{-1}\left(\int wydP\right),$$

assuming the existence of these quantities. Note that if these quantities exist for P, then they also exist for all P_t with $t \in \mathbb{R}$ and all g since K is a bounded function. Define the family of distributions as

$$\mathcal{P}_g := \{P_t : dP_t(w,y) = c(t)K(tg(w,y))dP(w,y), \text{ with } K(t) = 2(1+\exp(-2t))^{-1}, t \in \mathbb{R}\}.$$

Note that $K(0) = K'(0) = 1$. The Cramer-Rao lower bound for estimating $\psi(P)$ in the parametric family $\mathcal{P}_g$ is given by

$$\frac{1}{\mathbb{E}_P g^2}\left(\frac{\partial\psi(P_t)}{\partial t}\right)\left(\frac{\partial\psi(P_t)}{\partial t}\right)^\top\Bigg|_{t=0}.$$

For any matrix function A_t, we have

$$\frac{\partial A_t^{-1}}{\partial t} = -A_t^{-1}\left(\frac{\partial A_t}{\partial t}\right)A_t^{-1}.$$

For this family of distributions, the functional ψ is given by

$$\psi(P_t) = \left(\int ww^\top K(tg(w,y))dP\right)^{-1}\left(\int wyK(tg(w,y))dP\right).$$

This implies,

$$\begin{aligned}\frac{\partial\psi(P_t)}{\partial t} &= \left(\int ww^\top K(tg(w,y))dP\right)^{-1}\left(\int wyK'(tg(w,y))g(w,y)dP\right)\\ &- \left(\int ww^\top K(tg(w,y))dP\right)^{-1}\left(\int ww^\top K'(tg(w,y))g(w,y)dP\right)\times\\ &\qquad \left(\int ww^\top K(tg(w,y))dP\right)^{-1}\left(\int wydP\right).\end{aligned}$$

Define

$$\tilde{\psi}_P(w,y) := \left(\int ww^\top dP\right)^{-1} w(y - w^\top\alpha).$$

At $t = 0$, we get

$$\begin{aligned}\left.\frac{\partial\psi(P_t)}{\partial t}\right|_{t=0} &= -\left(\int ww^\top dP\right)^{-1}\left(\int ww^\top g(w,y)dP\right)\left(\int ww^\top dP\right)^{-1}\left(\int wydP\right)\\ &\quad + \left(\int ww^\top dP\right)^{-1}\left(\int wyg(w,y)dP\right)\\ &= \left(\int ww^\top dP\right)^{-1}\left(\int w(y-w^\top\alpha)g(w,y)dP\right) = \langle\tilde{\psi}_P, g\rangle_P,\end{aligned}$$

where $\langle g_1, g_2\rangle_P = \mathbb{E}_P(g_1 g_2)$. In a semi-parametric way, we are not restricted by a particular choice of g, and so the best semi-parametric lower bound on the variance for estimating $\psi(P)$ is lower bounded by the supremum over all $g \in L_2(P)$ such that $\mathbb{E}_P g = 0$. But it is easy to prove that

$$\sup_{g\in L_2(P):\,\mathbb{E}_P g=0}\frac{1}{\mathbb{E}_P g^2}\left(\frac{\partial\psi(P_t)}{\partial t}\right)\left(\frac{\partial\psi(P_t)}{\partial t}\right)^\top\Bigg|_{t=0} = \sup_{g\in L_2(P):\mathbb{E}_P g=0}\frac{\langle\tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P} = \mathbb{E}_P\tilde{\psi}_P\tilde{\psi}_P^\top.$$

It is interesting note that

$$\mathbb{E}_P\tilde{\psi}_P\tilde{\psi}_P^\top = \left(\mathbb{E}\left[WW^\top\right]\right)^{-1}\mathbb{E}\left[WW^\top(Y - W^\top\alpha)^2\right]\left(\mathbb{E}\left[WW^\top\right]\right)^{-1},$$

which is the asymptotic variance of the (normalized) least squares estimator $\hat{\alpha}$. Therefore, the ordinary least squares estimator is semi-parametrically efficient. The function $\tilde{\psi}_P$ is called the efficient influence function and the least squares estimator satisfies

$$\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}_P(W_i, Y_i) + o_p(1).$$

The proof of consistency of the sandwich estimator is included in the proof of semi-parametric efficiency of the sandwich estimator. The asymptotic variance of $\sqrt{n}(\hat{\alpha} - \alpha)$ is given by

$$\mathcal{V} := \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} \left(\mathbb{E}\left[WW^\top(Y - W^\top\alpha)^2\right]\right) \left(\mathbb{E}\left[WW^\top\right]\right)^{-1}.$$

As before, define the family of distributions as

$$\mathcal{P}_g := \{P_t : \, dP_t(w,y) = c(t)K(tg(w,y))dP(w,y), \text{ with } K(t) = 2(1{+}\exp(-2t))^{-1}, t \in \mathbb{R}\}.$$

The least squares slope functional at P_t is given by $\beta(P_t)$ which satisfies

$$\mathbb{E}\left[W(Y - W^\top\beta(P_t))K(tg(W,Y))\right] = 0 \quad \text{for all} \quad t \in \mathbb{R}.$$

Differentiating both sides with respect to t, we get

$$\mathbb{E}\left[W(Y - W^\top\beta(P_t))K'(tg(W,Y))g(W,Y)\right] - \mathbb{E}\left[WW^\top\frac{\partial\beta(P_t)}{\partial t}K(tg(W,Y))\right] = 0.$$

Therefore,

$$\frac{\partial\beta(P_t)}{\partial t} = \left(\mathbb{E}\left[WW^\top K(tg(W,Y))\right]\right)^{-1} \left(\mathbb{E}\left[W(Y - W^\top\beta(P_t))g(W,Y)\right]\right),$$

and at $t = 0$,

$$\left.\frac{\partial\beta(P_t)}{\partial t}\right|_{t=0} = \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} \left(\mathbb{E}\left[W(Y - W^\top\alpha)g(W,Y)\right]\right),$$

which was also what we derived above. Getting back to estimating the asymptotic variance $\mathcal{V}$ and proving efficiency first note that it is enough to provide efficiency bound for estimating $a^\top Va$ for any fixed vector a. This functional being scalar is easy to deal with than the matrix functional. The expressions for calculating the efficient influence function get more cumbersome in this case and so, it is easy to first prove a general result and then prove as a special case.

For any $t \in \mathbb{R}$ and given functionals $\phi_1(\cdot)$ and $\phi_2(\cdot)$, define the functional,

$$\psi_a(P_t) := a^\top \left[\phi_1(P_t)\right]^{-1} \phi_2(P_t) \left[\phi_1(P_t)\right]^{-1} a.$$

Suppose that $\tilde{\phi}_1(w,y)$ and $\tilde{\phi}_2(w,y)$ are the efficient influence functions for ϕ_1 and ϕ_2, that is,

$$\left.\frac{\partial\phi_1(P_t)}{\partial t}\right|_{t=0} = \langle\tilde{\phi}_1, g\rangle, \quad \text{and} \quad \left.\frac{\partial\phi_2(t)}{\partial t}\right|_{t=0} = \langle\tilde{\phi}_2, g\rangle.$$

By definition, $\mathbb{E}_P\tilde{\phi}_1 = \mathbb{E}_P\tilde{\phi}_2 = 0$. We now differentiate $\psi_a(P_t)$ with respect to t to get the efficient influence function for ψ_a.

$$\begin{aligned}\frac{\partial\psi_a(t)}{\partial t} &= 2a^\top \left[\phi_1(P_t)\right]^{-1} \left(\frac{\partial\phi_1(P_t)}{\partial t}\right) \left[\phi_1(P_t)\right]^{-1} \phi_2(P_t) \left[\phi_1(P_t)\right]^{-1} a \\ &\quad + a^\top \left[\phi_1(P_t)\right]^{-1} \left(\frac{\partial\phi_2(P_t)}{\partial t}\right) \left[\phi_1(P_t)\right]^{-1} a.\end{aligned}$$

Taking $t = 0$ and using efficient influence functions, we obtain

$$\left.\frac{\partial \psi_a(t)}{\partial t}\right|_{t=0}$$
$$= 2a^\top \left[\phi_1(P_t)\right]^{-1} \langle \tilde{\phi}_1, g\rangle \left[\phi_1(P_t)\right]^{-1} \phi_2(P_t) \left[\phi_1(P_t)\right]^{-1} a + a^\top \left[\phi_1(P_t)\right]^{-1} \langle \tilde{\phi}_2, g\rangle \left[\phi_1(P_t)\right]^{-1} a$$
$$= \left\langle 2a^\top \left[\phi_1(P_t)\right]^{-1} \tilde{\phi}_1 \left[\phi_1(P_t)\right]^{-1} \phi_2(P_t) \left[\phi_1(P_t)\right]^{-1} a + a^\top \left[\phi_1(P_t)\right]^{-1} \tilde{\phi}_2 \left[\phi_1(P_t)\right]^{-1} a, \quad g\right\rangle$$

Therefore, the efficient influence function for the functional $\psi_a(\cdot)$ is given by

$$\begin{aligned}\tilde{\psi}_a(w,y) = 2a^\top \left[\phi_1(P_t)\right]^{-1} \tilde{\phi}_1(w,y) \left[\phi_1(P_t)\right]^{-1} \phi_2(P_t) \left[\phi_1(P_t)\right]^{-1} a \\ + a^\top \left[\phi_1(P_t)\right]^{-1} \tilde{\phi}_2(w,y) \left[\phi_1(P_t)\right]^{-1} a.\end{aligned}$$

If T_1 and T_2 are semi-parametrically efficient estimators of ϕ_1 and ϕ_2 in the sense that

$$T_1 = \phi_1(P) + \frac{1}{n}\sum_{i=1}^n \tilde{\phi}_1(W_i, Y_i) + o_p(n^{-1/2}),$$
$$T_2 = \phi_2(P) + \frac{1}{n}\sum_{i=1}^n \tilde{\phi}_2(W_i, Y_i) + o_p(n^{-1/2}).$$

Then by formal expansion of inverse of matrices, we get

$$T_1^{-1} = \phi_1^{-1}(P) + \phi_1^{-1}(P)\frac{1}{n}\sum_{i=1}^n \tilde{\phi}_1(W_i, Y_i)\phi_1^{-1}(P) + o_p(n^{-1/2}).$$

This implies that the estimator $T_a = a^\top T_1^{-1} T_2 T_1^{-1} a$ satisfies the equation

$$T_a = \psi_a(P) + \frac{1}{n}\sum_{i=1}^n \tilde{\psi}_a(W_i, Y_i) + o_p(n^{-1/2}),$$

and thus T_a is a semi-parametrically efficient estimator for $\psi_a(P)$. Furthermore, this also implies that

$T_1^{-1} T_2 T_1^{-1}$ is a semi-parametrically efficient estimator for $\phi_1(P)^{-1}\phi_2(P)\phi_1^{-1}(P)$.

Getting back to $\mathcal{V}$, take $\phi_1(P_t) = c(t)\mathbb{E}\left[WW^\top K(tg(W,Y))\right]$ and

$$\phi_2(P_t) = c(t)\mathbb{E}\left[WW^\top (Y - W^\top \beta(P_t))^2 K(tg(W,Y))\right].$$

Here the expectation is taken with respect to $(W, Y) \sim P$. We will now compute the efficient influence function of ϕ_1 and ϕ_2.

$$\left.\frac{\partial \phi_1(P_t)}{\partial t}\right|_{t=0} = \mathbb{E}\left[WW^\top g(W,Y)\right] = \langle \tilde{\phi}_1, g\rangle, \quad \tilde{\phi}_1(w,y) = xx^\top - \mathbb{E}_P\left[WW^\top\right],$$

The sample mean estimator $T_1 = \sum_{i=1}^n W_i W_i^\top / n$ has this influence function and so is an efficient estimator.

Regarding the functional ϕ_2, we have

$$\begin{aligned}\frac{\partial \phi_2(P_t)}{\partial t} &= c'(t)\mathbb{E}\left[WW^\top (Y - W^\top \beta(P_t))^2 K(tg(W,Y))\right] \\ &\quad + c(t)\mathbb{E}\left[WW^\top (Y - W^\top \beta(P_t))^2 K'(tg(W,Y))g(W,Y)\right] \\ &\quad - 2c(t)\mathbb{E}\left[WW^\top (Y - W^\top \beta(P_t))W^\top \frac{\partial \beta(P_t)}{\partial t} K(tg(W,Y))\right] \\ \left.\frac{\partial \phi_2(P_t)}{\partial t}\right|_{t=0} &= \mathbb{E}\left[WW^\top (Y - W^\top \alpha)^2 g(W,Y)\right] \\ &\qquad - 2\mathbb{E}\left[WW^\top (Y - W^\top \alpha)W^\top\right] \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} \mathbb{E}[W(Y - W^\top \alpha)g(W,Y)] \\ &= \langle \tilde{\phi}_2, g \rangle,\end{aligned}$$

where $\tilde{\phi}_2(w,y) = \phi_2^\dagger(w,y) - \mathbb{E}_P \phi_2^\dagger(W,Y)$ and

$$\phi_2^\dagger(w,y) = xx^\top (y - w^\top \alpha)^2 - 2\mathbb{E}\left[WW^\top (Y - W^\top \alpha)W^\top\right] \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} w(y - w^\top \alpha).$$

Consider the usual estimator

$$T_2 = \frac{1}{n}\sum_{i=1}^n W_i W_i^\top (Y_i - W_i^\top \hat{\alpha})^2.$$

We already have the asymptotic linearity as

$$\hat{\alpha} = \alpha + \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} Z_n + o_p(n^{-1/2}), \quad Z_n := \frac{1}{n}\sum_{i=1}^n W_i(Y_i - W_i^\top \alpha) = O_p(n^{-1/2}).$$

Substituting this in T_2, we get

$$\begin{aligned}T_2 &= \frac{1}{n}\sum_{i=1}^n W_i W_i^\top \left(Y_i - W_i^\top \alpha - W_i^\top \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} Z_n\right)^2 + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n W_i W_i^\top (Y_i - W_i^\top \alpha)^2 - 2\frac{1}{n}\sum_{i=1}^n W_i W_i^\top (Y_i - W_i^\top \alpha) W_i^\top \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} Z_n + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n W_i W_i^\top (Y_i - W_i^\top \alpha)^2 \\ &\qquad - 2\mathbb{E}\left[WW^\top (Y - W^\top \alpha)W^\top\right] \left(\mathbb{E}\left[WW^\top\right]\right)^{-1} \frac{1}{n}\sum_{i=1}^n W_i(Y_i - W_i^\top \alpha) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n \phi_2^\dagger(W_i, Y_i) + o_p(n^{-1/2}).\end{aligned}$$

Therefore,

$$T_2 - \phi_2(P) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}_2(W_i, Y_i) + o_p(n^{-1/2}), \text{ and so is an efficient estimator.}$$

Finally, we arrive at the conclusion that the sandwich estimator of the asymptotic variance of the least squares estimator is a semi-parametrically efficient estimator of $\mathcal{V}$. Observe that T_1 and T_2 used above are same as $B_{n,M}$ and $V_{n,M}$ respectively. □

Bibliography

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Unpublished Manuscript*, pages 1–24.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1348–1356. Curran Associates, Inc.

Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045.