Supplementary material of "One-Step Regularized Estimator for High-Dimensional Regression Models"

Yi Wang, Donglin Zeng, Yuanjia Wang, Xingwei Tong

Beijing International Center for Mathematical Research, Peking University, China

Department of Biostatistics, Gillings School of Global Public Health,

University of North Carolina at Chapel Hill, U.S.A.

Department of Biostatistics, Mailman School of Public Health,

Columbia University, U.S.A.

School of Statistics, Beijing Normal University, China

Supplementary Material

S1 Connection between OSRE and Semiparametric Model

In a semiparametric model when P is indexed by a finite-dimension parameter, θ, and a nuisance parameter, η. Suppose that we are interested in the inference θ. In this case, m is the log-likelihood function and $v = \theta$. Following the semiparametric efficiency theory (Bickel et al. 1993, Chapter 3), h_n^* in our equation (2.1) is the least favorable direction for θ and $\nabla m[h_n^*] = I^{-1}(\theta_0)\dot{l}^*$, where I is the efficient information matrix and $\dot{l}^*$ is the efficient score function for θ. Therefore, our OSRE is equivalent to the one-step Newton-Raphson solution to the efficient score function.

In a fully nonparametric model, $v(P)$ is a functional of P. In this case, the efficient

influence function for $v(P)$ is ψ function satisfying

$$\dot{v}(P)(g) = \int \psi(Z; P) g dP,$$

where g is any $L_2(P)$ function with $\int g dP = 0$. Such a function exists and is unique (Bickel et al. (1993); Chapter 4). An initial estimator for $v(P)$ is $v(P_n)$. Thus, our OSRE beomes $v(P_n) - n^{-1} \sum \psi(Z_i, P_n)$. This is exactly the de-bias equation given in Kennedy (2022).

In summary, OSRE is equivalent to a one-step Newton-Raphson solution to the efficient score function in a semiparametric setting; and it reduces to the de-bias equation in Kennedy (2022) in a nonparametric setting.

S2 Proof of Theorem 1

In this section we will finish the proof of Theorem 1.

Proof. Supposing $d_{(n)}(\widehat{f}_n, f_{n0})$ converges to zero in probability, we have

$$\mathfrak{F}_n(\widehat{f}_n) - \mathfrak{F}_n(f_{n0}) = \left\langle v_n^*, \widehat{f}_n - f_{n0} \right\rangle_{(n)} + O_p\left(d_{(n)}^2(\widehat{f}_n, f_{n0})\right). \tag{S2.1}$$

As we have stated, our proposed estimator for θ_{n0} is defined as

$$\widetilde{\theta}_n = \widehat{\theta}_n - \mathbb{P}_n\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\}, \tag{S2.2}$$

where $\widehat{\theta}_n = \mathfrak{F}_n(\widehat{f}_n)$ is the plug-in estimator based on $\widehat{f}_n$.

By applying (S2.1), (S2.2) and Condition A.1, we have

$$\begin{aligned}
\widetilde{\theta}_n - \theta_{n0} &= \left(\mathfrak{F}_n(\widehat{f}_n) - \mathfrak{F}_n(f_{n0})\right) - \mathbb{P}_n \nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n] \\
&= \left\langle v_n^*, \widehat{f}_n - f_{n0} \right\rangle_{(n)} - (\mathbb{P}_n - P)\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\} \\
&\quad - P\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\} + O_p\left(d_{(n)}^2(\widehat{f}_n, f_{n0})\right).
\end{aligned}$$

With Conditions A.2 and A.3, we can further obtain

$$\begin{aligned}\widetilde{\theta}_n - \theta_{n0} =& P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, \widehat{f}_n - f_{n0}]\right\} - (\mathbb{P}_n - P)\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\} \\ & - P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[\widehat{h}_n, \widehat{f}_n - f_{n0}]\right\} + O_p\left(d^2_{(n)}(\widehat{f}_n, f_{n0})\right) \\ =& -(\mathbb{P}_n - P)\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\} - P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[\widehat{h}_n - h_n^*, \widehat{f}_n - f_{n0}]\right\} \\ & + O_p\left(d^2_{(n)}(\widehat{f}_n, f_{n0})\right).\end{aligned}$$

By the Cauchy-Schwarz inequality, Conditions A.4 and A.5 imply that

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[\widehat{h}_n - h_n^*, \widehat{f}_n - f_{n0}]\right\} = o_p(n^{-\frac{1}{2}}).$$

Consequently,

$$\begin{aligned}\sqrt{n}(\widetilde{\theta}_n - \theta_{n0}) =& -\sqrt{n}(\mathbb{P}_n - P)\left\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\right\} \\ & - \sqrt{n}P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[\widehat{h}_n - h_n^*, \widehat{f}_n - f_{n0}]\right\} + O_p\left(\sqrt{n}d^2_{(n)}(\widehat{f}_n, f_{n0})\right) \\ =& -\mathbb{G}_n\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\} + o_p(1)\end{aligned}$$

Finally, from Condition A.6,

$$-\mathbb{G}_n\{\nabla m(\boldsymbol{Z}, \widehat{f}_n)[\widehat{h}_n]\} = -\mathbb{G}_n\{\nabla m(\boldsymbol{Z}, f_{n0})[h_n^*]\} + o_p(1).$$

Then, we can conclude that

$$\sqrt{n}(\widetilde{\theta}_n - \theta_{n0}) = -\{\mathbb{G}_n \nabla m(\boldsymbol{Z}, f_{n0})[h_n^*]\} + o_p(1),$$

so we have proved the first part of the theorem. The second part of the theorem is due to the asymptotic linear expansion and Condition A.7. □

.

S3 Technical Conditions and proof of Theorem 3

In order to obtain the asymptotic properties for the OSRE listed in Theorem 3, we need the following technical assumptions:

B.1 $\boldsymbol{X}_i$ are i.i.d and there exists a constant U such that $\max_{i,j} |X_{i,j}| \leq U$.

B.2 The smallest eigenvalue $\Lambda^2_{\min}$ of Σ is larger than a constant $C_{\min}$. The largest eigenvalue $\Lambda^2_{\max}$ of Σ is smaller than a constant $C_{\max}$. Moreover, the diagonal elements of Σ are uniformly bounded by 1 after the normalization X's.

B.3 Define $C_0 = (32C_{\max}/C_{\min}) + 1$. We have $\rho(\Sigma, C_0 s_0) \leq \rho$, for some constant $\rho > 0$, where $\rho(A, k) = \max_{T \subseteq \{1,2,\ldots,p_n\}, |T| \leq k} \|A^{-1}_{T,T}\|_\infty$, where $|T|$ is the cardinality of T and $A_{T,T}$ denotes the block of A consisting of the rows and columns from T.

B.4 Let s_0 be the number of non-zero coefficients in $\boldsymbol{\beta}_{n0}$ and we assume $s_0 = O(n^{\alpha_0}/\log p_n)$, where $\alpha_0 < 1/2$.

B.5 Let s_Ω be the maximum sparsity level of the rows of $\Omega = \Sigma^{-1}$, which means $s_\Omega = \max_j \#\{j \neq k, \Omega_{j,k} \neq 0\}$. It holds that $s_\Omega = O(n^{\alpha_1}/\log p_n)$, where $\alpha_1 < 1/2$.

B.6 $\{\varepsilon_i\}_{i=1}^n$ are i.i.d with mean zero and variance σ^2_ε.

B.7 There exists a constant C such that $\|\boldsymbol{\beta}_{n0}\|_2 \leq C$ and $\|\boldsymbol{\beta}_{n0}\|_2 \neq 0$.

B.8 $\lim_{n\to\infty} 4\sigma^2_\varepsilon \boldsymbol{\beta}^{*T}_n \Omega_n \boldsymbol{\beta}^*_n \to c^2$.

Remark 1. Condition B.1 implies that covariates are uniformly bounded by a constant. In fact, most of variables are bounded and it is easy to find a uniform bound. Condition B.2 on eigenvalues of covariance matrix is common in high-dimensional models. [Van de Geer et al. (2014), Javanmard and Montanari (2014b), Javanmard et al. (2018)] Condition B.3 is also given by Javanmard and Montanari (2014b) to obtain a sharper bound on the bias of Lasso estimator. A large family of covariance matrices satisfy Condition B.3 , such as block diagonal matrices and circulant matrices, where $\Sigma_{i,j} = r^{|i-j|}$ for some $r \in (0, 1)$. Conditions B.4 and

B.5 control the sparsity of parameters and Σ^{-1}, similar conditions are also given by papers related to "de-biased" Lasso estimator [Van de Geer et al. (2014), Javanmard and Montanari (2014b) and Javanmard et al. (2018)]. Van de Geer et al. (2014) assumes $s_0 = O(n^{1/2}/\log p_n)$. Our method requires a slightly stronger assumption than Van de Geer et al. (2014) since our method applies to a much more general class of models are more flexible in applying to other models. These Conditions also imply that p_n can be larger than n and the largest p_n permitted is $o(\exp(n^{\min\{\alpha_1,\alpha_2\}}))$. Condition B.7 implies the real parameter vector can not be zero. Conditions B.6 and B.8 guarantee the asymptotic variance of OSRE tends to a constant.

Proof. Condition B.1 - B.3 has been verified in former statement. Using the same argument in last example, it is apparent that $\sqrt{n}d^2_{(n)}(\widehat{f}_n, f_{n0}) = o_p(1)$ which verifies Condition 4.

Consider

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[\widehat{h}_n - h^*_n, \widehat{h}_n - h^*_n]\right\} = 4(\widehat{\boldsymbol{\beta}}_n^T M - \boldsymbol{\beta}_n^{*T}\Omega)\Sigma(M^T\widehat{\boldsymbol{\beta}}_n - \Omega\boldsymbol{\beta}^*_n), \tag{S3.3}$$

where $M = \widehat{T}^{-2}\widehat{\Gamma}$. Note

$$\widehat{\boldsymbol{\beta}}_n^T M - \boldsymbol{\beta}_n^{*T}\Omega = \boldsymbol{\beta}_n^{*T}(M - \Omega) + (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)^T(M - \Omega) + (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)^T\Omega. \tag{S3.4}$$

Since Conditions B.1 - B.3 and B.5 hold, Van de Geer et al. (2014) shows that $\max_k \|M_k - \Omega_k\|_2 = O_p\left(\sqrt{s_\Omega \log p_n/n}\right)$, where M_k and Ω_k are k-th rows of M and Ω. Since $\|\boldsymbol{\beta}^*_n\|_2 \le C$, we have

$$\|\boldsymbol{\beta}_n^{*T}(M - \Omega)\|_2 = O_p\left(\sqrt{s_\Omega \log p_n/n}\right).$$

Thus, from the conditions that $s_\Omega = o(n^{\alpha_1}/\log p_n)$ for some $\alpha_1 < 1/2$ and $\Lambda_{\max} \le C_{\max}$, it gives $\sqrt{n}\boldsymbol{\beta}_n^{*T}(M - \Omega)\Sigma(M^T - \Omega)\boldsymbol{\beta}^*_n = o_p(1)$. Additionally, $s_0 = o(n^{\alpha_0}/\log p_n)$ for some $\alpha_0 < 1/2$ and $\Lambda_{\min} \ge C_{\min}$ imply that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_0)^T\Omega(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n) = o_p(1)$. Combing these

results with (S3.3) and (S3.4), we have proved that $\sqrt{n}d^2_{(n)}(\widehat{h}_n, h^*_n) = o_p(1)$. Condition 5 thus holds.

For Condition A.6, since

$$\nabla m(Z, f)[h] = -2(Y - \boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X}^T M^T\boldsymbol{\beta},$$

To verify Condition 6, recalling that for $h(\boldsymbol{X}) = -\boldsymbol{X}^T\boldsymbol{\gamma}$,

$$\nabla m(\boldsymbol{Z}, f)[h] = -(Y - \boldsymbol{X}^T\boldsymbol{\beta})\boldsymbol{X}^T\boldsymbol{\gamma}.$$

we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right) - \sqrt{n}P_{X,Y}\left\{\widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}(Y - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n)\right\}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) - \sqrt{n}\widehat{\boldsymbol{\beta}}_n^T M\Sigma(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\varepsilon_i\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\beta}_n^{*T}\Omega\Sigma(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)\right]\\
+&\sqrt{n}(\boldsymbol{\beta}_n^{*T}\Omega - \widehat{\boldsymbol{\beta}}_n^T M)\Sigma(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\varepsilon_i.
\end{aligned} \tag{S3.5}$$

For the first term, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\widehat{\boldsymbol{\beta}}_n^T M\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\beta}_n^{*T}(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)\right]\\
=&\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)^T M\widehat{\Sigma}_n(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) + \sqrt{n}\boldsymbol{\beta}_n^{*T}(M\widehat{\Sigma}_n - I)(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)\\
=&\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)^T(M\widehat{\Sigma}_n - I)(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n) + \sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n\|_2^2\\
+&\sqrt{n}\boldsymbol{\beta}_n^{*T}(M\widehat{\Sigma}_n - I)(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)
\end{aligned}$$

For $s_\Omega \lesssim n/\log p_n$, we have $\|M\widehat{\Sigma}_n - I\|_\infty \lesssim \sqrt{\log p_n/n}$. [Van de Geer et al. (2014)]As the result,

$$\left\|(M\widehat{\Sigma}_n - I)(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)\right\|_\infty = O_p\left(\sqrt{\log p_n/n}\sqrt{\frac{s_0\log p_n}{n}}\right).$$

Thanks to Condition B.4 holds, $\left\|(M\widehat{\Sigma}_n - I)(\boldsymbol{\beta}^*_n - \widehat{\boldsymbol{\beta}}_n)\right\|_\infty = o_p(n^{-1/2})$.

It has been proved that with probability tending to 1 [Van de Geer et al. (2014), Javanmard et al. (2018)],

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\|_2^2 \leq \frac{c_1 s_0 \sigma^2}{n} \log p_n, \tag{S3.6}$$

where c_1 is a constant.

According to (S3.6) and Condition B.4, we have

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\|_2^2 = o_p(n^{-\frac{1}{2}}). \tag{S3.7}$$

Combine former statement and Condition B.7, we can conclude

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\widehat{\boldsymbol{\beta}}_n^T M \boldsymbol{X}_i \boldsymbol{X}_i^T (\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\beta}_n^{*T} (\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) \right] = o_p(1).$$

As previous proof, we have

$$d_{(n)}^2(\widehat{h}_n, h_n^*) = (\widehat{\boldsymbol{\beta}}_n^T M - \boldsymbol{\beta}_n^{*T} \Omega) \Sigma (M^T \widehat{\boldsymbol{\beta}}_n - \Omega \boldsymbol{\beta}_n^*) = o(n^{-1/2}).$$

Combining with (S3.7), this implies

$$\sqrt{n} (\boldsymbol{\beta}_n^{*T} \Omega - \widehat{\boldsymbol{\beta}}_n^T M) \Sigma (\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) = o_p(1).$$

Therefore,

$$\begin{aligned} &\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\boldsymbol{\beta}}_n^T M \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_n \right) - \sqrt{n} 2 P_{X,Y} \left\{ \widehat{\boldsymbol{\beta}}_n^T M \boldsymbol{X}^T (Y - \boldsymbol{X}^T \widehat{\boldsymbol{\beta}}_n) \right\} \\ =& \frac{1}{\sqrt{n}} 2 \sum_{i=1}^{n} \widehat{\boldsymbol{\beta}}_n^T M \boldsymbol{X}_i \varepsilon_i + o_p(1). \end{aligned}$$

Thus, verifying Condition 6 is equivalent to verifying the asymptotic equicontinuity of the last term for the functional class $\{Q\boldsymbol{X}\varepsilon : \|Q - Q_n^*\|_1 \leq \delta_n\}$, where $Q_n^* = 2\boldsymbol{\beta}_{n0}^T \Omega$, $\delta_n = \delta^{-1} \max(s_0, s_\Omega) \sqrt{\log p_n / n}$.

To this end, let

$$G_{\delta_n} = \{\varepsilon \boldsymbol{X}^T \boldsymbol{\gamma}_1 - \varepsilon \boldsymbol{X}^T \boldsymbol{\gamma}_2 : d_{(n)}(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_n^*) \leq \delta_{1n}, \|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_n^*\|_1 \leq \delta_{2n}\},$$

where $\delta_{1n} = \delta^{-1}\sqrt{\max(s_\Omega, s_0)\log p_n/n}$, $\delta_{2n} = \delta^{-1}\max(s_\Omega, s_0)\sqrt{\log p_n/n}$, for some constant δ. By Markov's inequality and the symmetrization [Van der Vaart and Wellner (1996)], we have

$$P(\|\mathbb{G}_n\|_{G_{\delta_n}} > x) \leq \frac{2}{x} P \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i g(\boldsymbol{X}_i) \right\|_{G_{\delta_n}},$$

where ζ_i is Radmacher variable. Let $\mathcal{D} = \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : (\boldsymbol{X}^T\boldsymbol{\beta})^2 \leq 1\}$, which is an ellipsoid in $\mathbb{R}^{p_n}$. By Hoeffding's inequality [Van der Vaart and Wellner (1996)], the stochastic process is sub-Gaussian for the $l^2(\mathbb{P}_n)$-seminorm and we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} g^2(\boldsymbol{X}_i) &= \frac{1}{n}\sum_{i=1}^{n} \left[\varepsilon_i \boldsymbol{X}_i^T\boldsymbol{\gamma}_1 - \varepsilon_i \boldsymbol{X}_i^T\boldsymbol{\gamma}_2\right]^2 \\ &\leq \max_i (\boldsymbol{X}_i^T(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2))^2 \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.\end{aligned}$$

From the maximal inequality [Van der Vaart and Wellner (1996)], we conclude

$$\begin{aligned}&P_\varepsilon \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i g(\boldsymbol{X}_i) \right\|_{G_{\delta_n}} \\ \lesssim & P\left[\int_0^{\delta_n} \sqrt{\log N(\varepsilon, H_{\delta_n}, d_n)}d\varepsilon \left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)^{\frac{1}{2}}\right] \\ \lesssim & P\left[D^2(H_{\delta_n}, d_n)\right]^{\frac{1}{2}} \left(E\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)^{\frac{1}{2}}\end{aligned}$$

where $H_{\delta_n} = \{\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2 : d_{(n)}(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_n^*) \leq \delta_{1n}, \|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_n^*\|_1 \leq \delta_{2n}\}$ with the norm $d_n(\boldsymbol{\gamma}) = \max_i(\boldsymbol{X}_i^T\boldsymbol{\gamma})^2$l, and $D(H_{\delta_n}, d_n) = \int_0^\infty \sqrt{\log N(\varepsilon, H_{\delta_n}, d_n)}d\varepsilon$.

On the other hand, since $H_{\delta_n} \subseteq 2\delta_{1n}\mathcal{D}$ and $H_{\delta_n} \subseteq 2\delta_{2n}\mathcal{B}$, where $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : \|\boldsymbol{\beta}\|_1 \leq 1\}$, it suffices to bound $D(\mathcal{D}, d_n)$ and $D(\mathcal{B}, d_n)$ to complete the proof. If X is a random vector on $\mathbb{R}^{p_n}$, then A^*X is an isotropic random vector on $\mathbb{R}^{p_n}$. According to Lemma 4.4 of Bartlett et al. (2012), we have

$$D(\mathcal{B}, d_n) \leq cQh_n.$$

where $Q = \max_{1\leq i\leq n}\|\boldsymbol{X}_i\|_{l_\infty^{p_n}}$, $h_n = \log^{3/2} n \max\{\log^{1/2} n, \log^{1/2} p_n\}$. Moreover, Lemma 4.7

of Bartlett et al. (2012) implies that

$$P[D^2(\mathcal{D}, d_n)]^{\frac{1}{2}} \leq c_3 (EW^2)^{1/2} \sqrt{\log n} \log p_n,$$

where c_3 is an absolute constant and $W = \max \|A^* \boldsymbol{X}_i\|_2$. According to Condition 1, $Q \leq U$ and $W \leq U\sqrt{p_n}$. Hence,

$$P_\varepsilon \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i g(X_i) \right\|_{G_{\delta_n}} \lesssim \min\{\delta_{1n} \sqrt{p_n \log n} \log p_n, \delta_{2n} h_n\}.$$

Since $s_0 = O(n^{\alpha_0}/\log p_n)$ and $s_1 = O(n^{\alpha_1}/\log p_n)$, the right hand side tends to zero. This verifies the asymptotic equicontinuity in Condition 6 holds.

Combing with Conditions B.6 and B.8, we have finished the proof for Theorem 3. □

S4 Technical Conditions and proof of Theorem 4

In order to obtain the asymptotic properties of $\widehat{\theta}_n$, we need the following assumptions:

C.1 The number of nonzero components $q > 0$ is fixed and there is a constant $c_f > 0$ such that $\min_{1 \leq j \leq q} \|f_j\|_2 \geq c_f$.

C.2 The random variables ε_i are i.i.d with mean zero and $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$. Their tail probabilities satisfy $P(|\varepsilon_i| > x) \leq K \exp(-Cx^2), i = 1, \ldots, n$, for all $x \geq 0$ and for constants C and K.

C.3 $Pf_j(X_j) = 0$ and $f_j \in \mathrm{F}$, where F is the class of functions f on $[0, 1]$ whose kth derivative $f^{(k)}$ exists and satisfies a Lipschitz condition of order α:

$$\left| f^{(k)}(s) - f^{(k)}(t) \right| \leq C|s - t|^\alpha \text{ for } s, t \in [0, 1],$$

where $\alpha \in (0, 1]$ and let $d = k + \alpha$.

C.4 The covariate vector $\boldsymbol{X}$ has a continuous density and there exist constants C_1 and C_2 such that the density function g_j of X_j satisfies $0 < C_1 \leq g_j(x) \leq C_2 \leq \infty$ on $[0,1]$ for every $1 \leq j \leq p_n$.

C.5 In the projection $\pi[V_{1k}|\boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}]$, the number of nonzero components of $h_{jk}(X_j)$ functions $q_k > 0$ is fixed, h_{jk} is Lipschitz continuous with smooth parameter d and there is a constant $c_u > 0$ such that $\min_{2\leq j\leq q} \|u_{kj}\|_2 \geq c_u$, u_{kj} is the projection of $\boldsymbol{V}_{1,-k}$ into the space of X_j.

C.6 $\lambda_{n1} \asymp \sqrt{n\log(p_n m_n)}$ and $\widetilde{\lambda}_{kn1} \asymp \widetilde{\lambda}_{n1} \asymp \sqrt{n\log(p_n m_n)}$ uniformly in k and $m_n \asymp n^{1/(2d+1)}$ for $d > 3/2$.

C.7 Suppose that $\lambda_{n2} \leq O(n^{1/2})$ and satisfies

$$\frac{\lambda_{n2}}{n^(8d+3)/(8d+4)} = o(1),$$

and

$$\frac{n^{1/(4d+2)}\log^{1/2}(p_n m_n)}{\lambda_{n2}} = o(1).$$

C.8 Suppose that $\widetilde{\lambda}_{n2k} \asymp \widetilde{\lambda}_{n2} \leq O(n^{1/2})$ and satisfies

$$\frac{\widetilde{\lambda}_{n2}}{n^(8d+3)/(8d+4)} = o(1),$$

and

$$\frac{n^{1/(4d+2)}\log^{1/2}(p_n m_n)}{\widetilde{\lambda}_{n2}} = o(1).$$

C.9 Let $\boldsymbol{V}_{1,-k}^T \boldsymbol{\eta}_{k1}^* = \sum_{l\in A_{k1}} V_{1l}\eta_{k1l}^*$, where $A_{k1} = \{l : \eta_{k1l}^* \neq 0\}$ and

$$s_{nk}^*(X_{-1}) = \sum_{j\in A_{k2}} \sum_{l=1}^{m_n} \phi_l(X_j)\eta_{kjl}^*,$$

where $A_{k2} = \{j : s_{kj}^{*(n)}(x) \neq 0, \text{ for all } x\}$, and define

$$\boldsymbol{\eta}_{nk}^* = (\eta_{k1l_1}^*, \ldots, \eta_{k1l_{r_k}}^*, \eta_{kj_11}^*, \ldots, \eta_{kj_1m_n}^*, \ldots, \eta_{kj_{q_k}1}^*, \ldots, \eta_{kj_{q_k}m_n}^*),$$

where $l_u \in A_{k1}$ and $j_u \in A_{k2}$. Suppose $\boldsymbol{\eta}^*_{nk1}$ and s^*_{nk} maximize $P[V_{1k} - \boldsymbol{V}^T_{1,-k}\boldsymbol{\eta}^*_{nk1} - s(X_{-1})]^2$ in the sieve space which is the linear space of B-spline functions in this example. Besides, suppose constant Q such that $q_k < Q$ and it satisfies $\|\boldsymbol{\eta}^*_{nk}\|_2 \leq C_\eta$ for all k and a constant C_η.

C.10 Suppose

$$\lim_{n\to\infty} \boldsymbol{\kappa}^{*T}_{n1}\Omega_{n11}\boldsymbol{\kappa}^*_{n1} = c^2,$$

where $\boldsymbol{\kappa}^*_{n1} = (\kappa^*_{n11}, \ldots, \kappa^*_{n1m_n})$, $\kappa^*_{n1k} = \int f_{01}(t)\phi_k(t)dt$, Ω_{n11} is the first m_n rows and lines of Ω_n which is the inverse of Covariance matrix of B-spline functions of $\{X_1, \ldots, X_{p_n}\}$.

Remark 2. Conditions C.1, C.3 and C.4 are standard conditions for nonparametric additive models. These conditions are needed to estimate the nonzero additive components at optimal rate, even if q important variables are known. Condition C.1 can be slightly relaxed to q increase in $\log n$ scale. Condition C.2 strengthens the assumptions needed for nonparametric estimation of a nonparametric additive model. Condition C.6 are the scale of λ_{n1} which is suggested by Huang et al. (2010) to guarantee Lasso estimators in the first step are well enough to be used as weights in the adaptive group Lasso, and the same scale of $\widetilde{\lambda}_{n1}$ also be applied to the group Lasso of $\pi[V_{1k}|\boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}]$. Condition C.7 is the condition to obtain converge rate of the function. Conditions C.1 - C.4 and C.7 are also given by Huang et al. (2010). Condition C.5 and C.9 ensures the sparse structure of the projection $\pi[V_{1k}|\boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}]$. Since basis functions of B-spline are almost orthogonal to each other and most of covariates are independent from each other, $\pi[V_{1k}|\boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}]$ can satisfy the sparse structure. At least conditions C.5 and C.9 hold if all the covariates are independent. Condition C.9 also implies $\|\boldsymbol{\eta}_{nk0}\|_2$ won't tend to ∞. Besides, Condition C.8 is the same condition as C.7 which is applied to $\pi[V_{1k}|\boldsymbol{V}_{1,-k}, X_2, \ldots, X_{p_n}]$. Besides, Conditions

C.7 and C.8 allow $\lambda_{n2} = O(n^{1/2})$ and $\widetilde{\lambda}_{n2} = O(n^{1/2})$. Then, we can obtain p_n is at most $o(\exp(n^{2d/(2d+1)}))$ by substituting λ_{n2} and $\widetilde{\lambda}_{n2}$ into the two equations in Conditions C.7 and C.8. Condition C.10 guarantee the asymptotic variance of the estimator tends to a constant.

Proof. Conditions A.1 - A.3 have been verified in the main text. The remaining conditions need to verify are A.4 - A.6.

For Condition A.4, since Conditions C.1 - C.4, C.6 and C.7 hold, it was proved by Huang et al. (2010) that

$$\sum_{j=1}^{p_n} \|\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n\|_2^2 = O_p\left(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right) \tag{S4.8}$$

and

$$\sum_{j=1}^{p_n} \|\widehat{f}_{nj} - f_{n0j}\|_2^2 = O_p(\frac{m_n}{n} + \frac{1}{n} + \frac{1}{m_n^{2d}} + \frac{4m_n\lambda_{n2}^2}{n^2}). \tag{S4.9}$$

Thus since $\lambda_{n1} = O(\sqrt{n\log(p_n m_n)})$, $m_n = O(n^{1/(2d+1)})$ and $\lambda_{n2} \leq O(n^{1/2})$,

$$\sqrt{n}d_{(n)}^2(\widehat{f}_n, f_{n0}) = O_p(n^{1/2}n^{-2d/(2d+1)}) = O_p(n^{(1/2-d)/(2d+1)}).$$

Since $d > 3/2$, $d_{(n)}(\widehat{f}_n, f_{n0}) = o_p(n^{-1/4})$.

To verify Condition A.5, we consider $< \widehat{h}_n - h_n^*, \widehat{h}_n - h_n^* >_{(n)}$. Since $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)^T$ has i.i.d bounded rows and $\widetilde{\lambda}_{nk2}$ has the same order of $\widetilde{\lambda}_{n2}$ in Condition C.8, using the proof of (S4.8), we have

$$\sum_{j=2}^{p_n} \|\widehat{\boldsymbol{\eta}}_{kj} - \boldsymbol{\eta}_{nkj}^*\|_2^2 = O_p(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}). \tag{S4.10}$$

Then we have

$$\begin{aligned} |\widehat{\tau}_k^2 - \tau_k^2| =& |\boldsymbol{\xi}_k^T\boldsymbol{\xi}_k/n - \tau_k^2\| + |\boldsymbol{\xi}_k^T V_{A_k}(\widehat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{nk}^*)/n| \\ &+ |\boldsymbol{\xi}_k^T V_{A_k}\boldsymbol{\eta}_{nk}^*/n| + |\boldsymbol{\eta}_{nk}^{*T}V_{A_k}^T V_{A_k}(\widehat{\boldsymbol{\eta}}_{kn} - \boldsymbol{\eta}_{nk}^*)/n|, \end{aligned} \tag{S4.11}$$

where $\boldsymbol{\xi}_k = (\xi_{k1}, \ldots, \xi_{kn})^T$ with $\xi_{ki} = V_{1k} - s_{nk}^*(X_{i,A_k})$. Recalling the definition in condition C.9, V_{A_k} is obtained by removing all m_n columns that are not in the set A_k of $\boldsymbol{V}$.

We now consider each term in the right-hand side of (S4.11). Without loss of generally, we suppose the first q_k components of $s_k^{*(n)}$ are not zero and the others are zero. For the first term, we have

$$|\boldsymbol{\xi}_k^T\boldsymbol{\xi}_k/n - \tau_k^2| \leq 2(|\boldsymbol{\epsilon}_k^T\boldsymbol{\epsilon}_k/n - \tau_k^2| + |\boldsymbol{\zeta}_{kn}^T\boldsymbol{\zeta}_{kn}/n|) \lesssim n^{-1/2} + qm_n^{-2d},$$

where $\boldsymbol{\zeta}_{kn} = (\zeta_{1kn}, \ldots, \zeta_{nkn})^T$ with $\zeta_{ikn} = \sum_{j\in A_k}(s_k^{*(n)}(X_{ij}) - s_{nk}^*(X_{ij}))$. The last step holds for

$$\|s_k^{*(n)}(X_{ij}) - s_{nk}^*(X_{ij})\|_2 = O\left(m_n^{-d}\right).$$

For the second term in (S4.11),

$$|\boldsymbol{\xi}_k^T V_{A_k}(\widehat{\boldsymbol{\eta}}_{kn} - \boldsymbol{\eta}_{nk0})/n| \leq \|\boldsymbol{\xi}_k^{*T}\|_2 \|V_{A_k}(\widehat{\boldsymbol{\eta}}_{kn} - \boldsymbol{\eta}_{nk0})/n\|_2,$$

where $\boldsymbol{\xi}_k^* = V_{A_k}(V_{A_k}^T V_{A_k})^{-1}V_{A_k}^T\boldsymbol{\xi}_k$. By equations (36) and (37) of Huang et al. (2010),

$$\|\boldsymbol{\xi}_k^*\|_2^2 = O_p(qm_n + 1 + qnm_n^{-2d}). \tag{S4.12}$$

Lemma 3 of Huang et al. (2010) implies there exist constants c_1 and c_2 such that

$$c_1 m_n^{-1} \leq \sigma_{\min}(V_{A_k}^T V_{A_k}/n) \leq \sigma_{\max}(V_{A_k}^T V_{A_k}/n) \leq c_2 m_n^{-1}. \tag{S4.13}$$

Combing previous results, we have

$$\|V_{A_k}(\widehat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{nk}^*)/n\|_2^2 \lesssim n^{-1}m_n^{-1}\left(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right).$$

Consequently, we have

$$\begin{aligned}
&|\boldsymbol{\xi}_k^T V_{A_k}(\widehat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{nk}^*)/n|^2 \\
=&O_p\left(n^{-1}(qm_n + 1 + qnm_n^{-2d})\left(\frac{m_n}{n} + \frac{1}{n} + \frac{1}{m_n^{2d}} + \frac{4m_n\lambda_{n2}^2}{n^2}\right)\right).
\end{aligned}$$

For the third term in (S4.11), from (S4.12), (S4.13) and Condition C.9, we obtain

$$|\boldsymbol{\xi}_k^T V_{A_k}\boldsymbol{\eta}_{nk}^*/n|^2 \lesssim n^{-1}m_n^{-1}(qm_n + 1 + qnm_n^{-2d}).$$

Using (S4.10), (S4.13) and Condition C.9, we have

$$|\boldsymbol{\eta}_{nk}^{*T}V_{A_k}^T V_{A_k}(\widehat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{nk}^*)/n|^2 \lesssim m_n^{-2}\left(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right).$$

Combining the above results, since $m_n = O(n^{1/(2d+1)})$ and $\lambda_{n2} \leq O(n^{1/2})$, for $d > 3/2$, we conclude $|\widehat{\tau}_k^2 - \tau_k^2| = o_p(n^{-1/4})$. Since $1/\tau_k^2 = O(1)$, this also implies$1/\widehat{\tau}_k^2 - 1/\tau_k^2 = o_p(n^{-1/4})$.

With this result, let $\widehat{M} = \widehat{T}^{-2}\widehat{C}$ so

$$\begin{aligned}\|\widehat{M}_k - M_k\|_2 &\leq \|\widehat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_{nk}^*\|/\widehat{\tau}_k^2 + \|\boldsymbol{\eta}_{nk}^*\|_2(1/\widehat{\tau}_k^2 - 1/\tau_k^2) \\ &= o_p(n^{-1/4}).\end{aligned} \tag{S4.14}$$

Clearly,

$$\|\widehat{\boldsymbol{\kappa}}^T\widehat{M} - \boldsymbol{\kappa}^T M\|_2^2 \leq \|(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa})^T M\|_2^2 + \|(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa})^T(\widehat{M} - M)\|_2^2 + \|\boldsymbol{\kappa}^T(\widehat{M} - M)\|_2^2.$$

Define $\Sigma_{V_k} = P[\boldsymbol{V}_{i,A_k}\boldsymbol{V}_{i,A_k}^T]$, where $\boldsymbol{V}_{i,A_k}$ is obtained by removing all components that are included in A_{k1} and all m_n components that are not in the set A_k of $\boldsymbol{V}_i$. By taking expectation in the proof of (S4.13), there exist constants c_3 and c_4 such that

$$c_3 m_n^{-1} \leq \sigma_{\min}(\Sigma_{V_k}) \leq \sigma_{\max}(\Sigma_{V_k}) \leq c_4 m_n^{-1}. \tag{S4.15}$$

Since M is first m_n columns of $\Sigma_{V_k}^{-1}$, it holds

$$\|(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa})^T M\|_2^2 \lesssim m_n\left(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4m_n^2\lambda_{n2}^2}{n^2}\right).$$

Under Condition C.3, C.6 and C.8, we obtain $\|(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa})^T M\|_2^2 = o_p(m_n n^{-1/2})$. Thus, combing with (S4.14), we have

$$\|\widehat{\boldsymbol{\kappa}}^T\widehat{M} - \boldsymbol{\kappa}^T M\|_2^2 = o_p(m_n n^{-1/2}). \tag{S4.16}$$

By the properties of spline [De Boor et al. (1978)], there exists a positive constant c such that

$$d_{(n)}^2(\widehat{h}_n, h_n^*) \leq c m_n^{-1}\|\widehat{\boldsymbol{\kappa}}\widehat{M} - \boldsymbol{\kappa}M\|_2^2.$$

This implies $d^2_{(n)}(\widehat{h}_n, h^*_n) = o_p(n^{-1/2})$. We have verified Condition A.5.

To verify Condition A.6, recall that

$$\nabla m(\boldsymbol{Z}, \hat{f})[\hat{h}] = -\widehat{\boldsymbol{\kappa}}^T \widehat{M} V (Y - V^T \widehat{\boldsymbol{\beta}}).$$

Therefore, similar decomposition in (S3.5) can be applied, then we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\boldsymbol{\kappa}}^T \widehat{M} V_i \left(Y_i - V_i^T \widehat{\boldsymbol{\beta}}_n \right) - \sqrt{n} P_{\boldsymbol{V},Y} \left(\widehat{\boldsymbol{\kappa}}^T \widehat{M} \boldsymbol{V} \left(Y - \boldsymbol{V}^T \widehat{\boldsymbol{\beta}}_n \right) \right) \\
=&\sqrt{n} \left[(\widehat{\boldsymbol{\kappa}}^T \widehat{M} - \boldsymbol{\kappa}^T M) \widehat{\Sigma}_{V_k} (\boldsymbol{\beta}_{n0} - \widehat{\boldsymbol{\beta}}_n) \right] + \sqrt{n} \left[\boldsymbol{\kappa}^{*T} \left(\Sigma_{V_k}^{-1} \widehat{\Sigma}_{V_k} - I \right) (\boldsymbol{\beta}_{n0} - \widehat{\boldsymbol{\beta}}_n) \right] \\
+&\sqrt{n} \left[(\widehat{\boldsymbol{\kappa}}^T \widehat{M} - \boldsymbol{\kappa}^T M) \Sigma_{V_k} (\boldsymbol{\beta}_{n0} - \widehat{\boldsymbol{\beta}}_n) \right] + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - V_i^T) V_i^T \widehat{\boldsymbol{\kappa}}^T \widehat{M},
\end{aligned}$$

where $\widehat{\Sigma}_{V_k} = V_k^T V_k / n$, $\boldsymbol{\kappa}^{*T} = (\boldsymbol{\kappa}^T, 0, \ldots, 0)$, $\boldsymbol{\beta}_{n0}$ maximizes m-function in the sieve space. As for the first term, combining (S4.8) with (S4.13) and (S4.16), we have

$$\sqrt{n} \left[(\widehat{\boldsymbol{\kappa}}^T \widehat{M} - \boldsymbol{\kappa}^T M) \widehat{\Sigma}_{V_k} (\boldsymbol{\beta}_{n0} - \widehat{\boldsymbol{\beta}}_n) \right] = O_p \left(n^{1/4} \left(\frac{m_n^2}{n} + \frac{m_n}{n} + \frac{1}{m_n^{2d-1}} + \frac{4 m_n^2 \lambda_{n2}^2}{n^2} \right)^{1/2} \right).$$

Since Condition 6 holds, we know that the first term is $o_p(1)$. Follow the proof of Theorem 7 of Javanmard and Montanari (2014a) by replacing the $C_{\min} \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq C_{\max}$ with (S4.15), we have

$$\| \Sigma_{V_k}^{-1} \widehat{\Sigma}_{V_k} - I \|_\infty = O_p \left(\sqrt{\frac{\log p_n m_n}{n}} \right).$$

Combine with (S4.8), we have

$$\sqrt{n} \left[\boldsymbol{\kappa}^{*T} \left(\Sigma_{V_k}^{-1} \widehat{\Sigma}_{V_k} - I \right) (\boldsymbol{\beta}_{n0} - \widehat{\boldsymbol{\beta}}_n) \right] = o_p(1).$$

It follows from (S4.8), (S4.15) and (S4.16) that the third term is $o_p(1)$. Thus, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\boldsymbol{\kappa}}^T \widehat{M} V_i \left(Y_i - V_i^T \widehat{\boldsymbol{\beta}}_n \right) - \sqrt{n} P_{\boldsymbol{V},Y} \left(\widehat{\boldsymbol{\kappa}}^T \widehat{M} \boldsymbol{V} \left(Y - \boldsymbol{V}^T \widehat{\boldsymbol{\beta}}_n \right) \right) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\boldsymbol{\kappa}}^T \widehat{M} V_i \nu_i + o_p(1).
\end{aligned}$$

where $\nu_i = Y_i - V_i^T \boldsymbol{\beta}_{n0}$ Thus, verifying Condition A.6 is equivalent to proving the equicontinuity of the right-hand side of the above expression.

To this end, let

$$G_{\delta_n} = \{\nu V^T \boldsymbol{\gamma}_1 - \nu V^T \boldsymbol{\gamma}_2 : d^2_{(n)}(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}^*_n) \leq \delta_{1n}\},$$

where $\boldsymbol{\gamma}^*_n = \kappa^T M$, $\delta_{1n} = \delta^{-1} m_n n^{-1/2}$, for some constant δ. Following to the proof of Theorem 3, we can conclude

$$P[D^2(\mathcal{D}, d_n)]^{\frac{1}{2}} \leq c(EW^2)^{1/2} \sqrt{\log n} \log p_n m_n,$$

Since B-spline basis functions are bounded by a constant U According to Condition C.5 and C.9, each line of Σ_V^{-1} has at most qm_n columns of non-zero components. Recalling the definition of W, $W \leq U\sqrt{qm_n}$. Hence,

$$P_\nu \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nu_i g(X_i) \right\|_{G_{\delta_n}} \lesssim \sqrt{\delta_{1n}} \sqrt{m_n \log n} \log p_n m_n.$$

Since Condition C.6 holds, the right hand side tends to zero. Then, we can verify the asymptotic equicontinuity in Condition A.6. Finally, Condition A.7 follows naturally from condition C.10.

□

S5 OSRE for coefficient inference in Lasso

Single coefficient example has been widely learned in recent years, which is called as "de-biased" estimator. In this section we will use it as another example. Consider n i.i.d samples $(\boldsymbol{X}_i, Y_i)$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_n})^T \in \mathbb{R}^{p_n}$, where one of $X's$ is one and the others have mean zero for $k > 1$. Moreover, it holds

$$Y_i = \sum_{j=1}^{p_n} X_{ij}\beta_{n0j} + \varepsilon_i,\ P[\varepsilon_i|\boldsymbol{X}_i] = 0, \tag{S5.1}$$

where $\boldsymbol{\beta}_{n0} = (\beta_{n01}, \ldots, \beta_{n0p_n})^T$ is the vector of parameters and ε_i is a random variable representing the noise in the i-th response variable. Let $\mathcal{F}_n = \{f(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\beta_j\}$, which is a linear functional class. In addition, we equip $\mathcal{F}_n$ with an inner product as

$$< f_1, f_2 >_{(n)} = P\left\{f_1(\boldsymbol{X})f_2(\boldsymbol{X})\right\} = \sum_{j,k=1}^{p_n} \beta_j P[X_jX_k]\beta_k \ \text{ for all } f_1, f_2 \in \mathcal{F}_n.$$

The true function $f_{n0}(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\beta_{n0j} \in \mathcal{F}_n$ is assumed to be the unique maximizer for $P\{m(\boldsymbol{X}, Y, f)\}$ with $m(\boldsymbol{X}, Y, f) = -(Y - f(\boldsymbol{X}))^2/2$. Clearly,

$$P\left\{\nabla m(\boldsymbol{X}_i, Y_i, f_{n0})[h]\right\} = P\left\{\left(Y_i - f_{n0}(\boldsymbol{X}_i)\right)h(\boldsymbol{X}_i)\right\},$$

and

$$P\left\{\nabla^2 m(\boldsymbol{X}_i, Y_i, f_{n0})[h_1, h_2]\right\} = -P\left\{h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\right\}.$$

Suppose that we are interested in the first component of $\boldsymbol{\beta}_{n0}$, . Then $\mathfrak{F}_n(f) = \beta_{n01}$ which is equivalent to $P[\boldsymbol{e}_1^T\Sigma_n^{-1}\boldsymbol{X}f(\boldsymbol{X})]$, where $\boldsymbol{e}_1$ is a p_n-vector with only the first element is one and the others are zero, and Σ_n is the covariance matrix of $\boldsymbol{X}$ assumed to be non-singular. Furthermore, for $h_n(\boldsymbol{X}) = \sum_{j=1}^{p_n} X_j\gamma_j$, it holds $\nabla\mathfrak{F}_n(f_{n0})[h_n] = \gamma_1$. To construct OSRE for β_{n01}, the key step is to obtain h_n^* as given in Condition 3. For the linear model, we show

$$h_n^*(\boldsymbol{X}) = -(P\left\{g_n^2(\boldsymbol{X})\right\})^{-1}g_n(\boldsymbol{X}), \tag{S5.2}$$

where $g_n(\boldsymbol{X}) = X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n})$, and $\pi(X_1|X_2, \ldots, X_{p_n})$ is the $L^2(P)$ projection of X_1 onto the linear span of $X_2, \ldots, X_{p_n}$.

To see this, since $\pi(X_1|X_2, \ldots, X_{p_n})$ is the $L^2(P)$ projection of X_1 onto the linear span of $X_2, \ldots, X_{p_n}$, we obtain

$$P\left\{(X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n}))\pi(X_1|X_2, X_3, \ldots, X_{p_n})\right\} = 0,$$

and

$$P\left\{(X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n}))X_k\right\} = 0 \text{ , for all } k = 2, 3, \ldots, p_n.$$

Thus,

$$\begin{aligned} P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, X_1]\right\} =& (Pg_n^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi[X_1|X_2, X_3, \ldots, X_{p_n}])X_1] \\ =& (Pg_n^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n}))^2] \\ =& 1 \end{aligned}$$

and for $k > 1$,

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, X_k]\right\} = (Pg_n^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi[X_1|X_2, X_3, \ldots, X_{p_n}])X_k] = 0.$$

Consequently, for any $h_n \in \mathcal{F}_n$ with $h_n(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j \gamma_j$, we obtain

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, h_n]\right\} = \gamma_1 = \nabla \mathfrak{F}(f_{n0})[h_n].$$

In other words, h_n^* satisfies Condition 3.

Therefore, suppose $\widehat{\boldsymbol{\beta}}_n$ is an initial estimator of $\boldsymbol{\beta}_n^*$ and we can find a proper estimator for h_n^*, denoted by $\widehat{h}_n$. The OSRE for β_{n1}^* is then given as

$$\widetilde{\beta}_{n1} = \widehat{\beta}_{n1} - \frac{1}{n} \sum_{i=1}^{n} \widehat{h}_n(\boldsymbol{X}_i) \left(Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_n\right), \tag{S5.3}$$

where $\widehat{\beta}_{n1}$ is the first coordinate of $\widehat{\boldsymbol{\beta}}_n$.

Without loss of generality, we suppose the parameter we are interested in is the first coordinate of linear parameter. Suppose the linear regression model is defined as (S5.1). The vector parameter $\boldsymbol{\beta}_{n0}$ is sparse, then the initial estimator $\widehat{\boldsymbol{\beta}}_n$ in (S5.3) can be estimated using the Lasso method:

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p_n}} \left\{ \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{S5.4}$$

where $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ is the l_1-norm on $\mathbb{R}^{p_n}$ and $\lambda \geq 0$ is a penalty parameter. Obviously, the Lasso estimator of f_{n0} is

$$\widehat{f}_n(\boldsymbol{X}_i) = \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n, \tag{S5.5}$$

where $\widehat{\boldsymbol{\beta}}_n$ is defined by (S5.4).

Next, we should estimate h_n^* defined by (S5.2). Recalling the definition of h_n^*, we should first estimate $\pi(X_1|X_2, X_3, \ldots, X_{p_n})$, the projection of X_1 on the linear space spanned by $X_1, ..., X_{p_n}$. This estimation can be treated as another high-dimensional linear regression problem, so we adopt Lasso to estimate the coefficients:

$$\widehat{\boldsymbol{\eta}}_1 = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2n}\|\boldsymbol{X}_1 - \boldsymbol{X}_{-1}\boldsymbol{\eta}\|_2^2 + \widetilde{\lambda}\|\boldsymbol{\eta}\|_1 \right\},$$

where $\widehat{\boldsymbol{\eta}}_1 = (\widehat{\eta}_{11}, \ldots, \widehat{\eta}_{1p_n})$, $\boldsymbol{X}_1 = (X_{i1}, \ldots, X_{n1})^T$, $\boldsymbol{X}_{-1}$ is the sub-matrix of $\boldsymbol{X}$ obtained by removing the first column. We then obtain

$$\widehat{g}_n(\boldsymbol{X}) = X_1 - \widehat{\pi}(X_1|X_2, \ldots, X_{p_n}) = X_1 - X_{-1}^T\widehat{\boldsymbol{\eta}}_1,$$

where $X_{-1} = (X_2, \ldots, X_{p_n})^T$. Next, we estimate $Pg_n^2(X)$ using

$$\widehat{\tau}_1^2 = \|\boldsymbol{X}_1 - \boldsymbol{X}_{-1}^T\widehat{\boldsymbol{\eta}}_1\|_2^2/n + \lambda\|\widehat{\boldsymbol{\eta}}_1\|_1.$$

Consequently, the estimator for h_n^* is given as

$$\widehat{h}_n(\boldsymbol{X}) = -\widehat{g}_n(\boldsymbol{X})/\widehat{\tau}_1^2. \tag{S5.6}$$

Since the second part of (S2.2) is

$$\mathbb{P}_n\left\{\nabla m(Z,\widehat{f}_n)[\widehat{h}_n]\right\} = \mathbb{P}_n\left\{(Y-\widehat{f}_n(X))\widehat{h}_n(X)\right\},$$

by applying (S5.5) and (S5.6) to (S2.2), the one-step regularized estimator for β_{n01} is

$$\widetilde{\beta}_{n1} = \widehat{\beta}_{n1} + \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_1^{-2}\left(X_{i1} - \sum_{j=2}^{p_n} X_{ij}\widehat{\eta}_{1j}\right)\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right), \tag{S5.7}$$

where $\widehat{\beta}_{n1}$ is the first element of the initial Lasso estimator $\widehat{\boldsymbol{\beta}}_n$. As a note, the OSRE is exactly the same as the de-biased Lasso estimator in Van de Geer et al. (2014) and Javanmard et al. (2018).

To state the asymptotic properties for the OSRE, we need the following assumptions:

B.9 Let s_1 be the number of non-zero elements of the first row of $\Omega = \Sigma^{-1}$. We assume $s_1 = O(n^{\alpha_1}/\log p_n)$, where $\alpha_1 < 1/2$.

B.10 Let Ω_{11n} be the first line and first row element of Ω_n, suppose $\lim_{n\to\infty}\Omega_{11n} = \Omega_{11}$.

Remark 3. Condition B.9 controls the sparsity of parameters and Σ^{-1}, similar conditions are also given by papers related to "de-biased" Lasso estimator [Van de Geer et al. (2014), Javanmard and Montanari (2014b) and Javanmard et al. (2018)]. These Conditions also imply that p_n can be larger than n and the largest p_n permitted is $o(\exp(n^{\min\{\alpha_1,\alpha_2\}}))$. Conditions B.6 and B.10 guarantee the asymptotic variance of OSRE tends to a constant.

Theorem 1. *Suppose that Conditions B.1 - B.4, B.6, B.9,B.10 hold true. Furthermore,* $\lambda \asymp \sqrt{\log p_n/n}$ *in* (S5.4) *and* $\widetilde{\lambda} \asymp \sqrt{\log p_n/n}$ *in* (S5). *Then*

$$\sqrt{n}(\widetilde{\beta}_{n1} - \beta_{n01}) \xrightarrow{p} N(0, c^2),$$

where $c^2 = \sigma_\varepsilon^2\Omega_{11}$, .

Proof. In order to prove this theorem, we need to verify all the conditions we have listed before. Conditions A.1 - A.3 have been verified in Section 2.3. Thus, the remaining conditions to verify are Conditions A.4 - A.7. First, let us consider $d_{(n)}(\widehat{f}_n, f_{n0})$. Since (S3.6) and Condition B.2 holds, we have

$$d^2_{(n)}(\widehat{f}_n, f_{n0}) = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)^T \Sigma (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n) = O_p\left(\frac{s_0 \log p_n}{n}\right).$$

Thus, Condition B.4 implies $\sqrt{n} d^2_{(n)}(\widehat{f}_n, f_{n0}) = o_p(1)$ so Condition 4 holds.

We now consider

$$P\left\{\nabla^2 m(Z, f_{n0})[\widehat{h}_n - h^*_n, \widehat{f}_n - f_{n0}]\right\} = (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^*_n)^T \Sigma (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n),$$

where $\boldsymbol{\gamma}^*_n = (1, \boldsymbol{\eta}^T_{n01})^T / \tau^2_{10}$, $\boldsymbol{\eta}_{n01}$ is the vector of coefficient of $\pi(X_1 | X_2, \ldots, X_{p_n})$, $\tau^2_{10} = P g^2_n(\boldsymbol{X})$ and $\widehat{\boldsymbol{\gamma}}_n = (1, \widehat{\boldsymbol{\eta}}^T_1)^T / \widehat{\tau}^2_1$. Since Conditions B.1, B.2 and B.9 hold and $\widetilde{\lambda} \asymp \sqrt{\log p_n / n}$, as proved by Theorem 2.4 of Van de Geer et al. (2014) that $\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^*_n\|_2 = O_p\left(\sqrt{s_1 \log p_n / n}\right)$, we have

$$\sqrt{n} |(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^*_n)^T \Sigma (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)| = O_p\left(\frac{\sqrt{s_0 s_1} \log p_n}{\sqrt{n}}\right). \tag{S5.8}$$

We use $s_0 = O(n^{\alpha_0} / \log p_n)$ and $s_1 = O(n^{\alpha_1} / \log p_n)$ to obtain that $\sqrt{n} |(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}^*_n)^T \Sigma (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n)| = o_p(1)$. This verifies Condition A.5.

To verify Condition A.6, recalling that for $h(\boldsymbol{X}) = -\boldsymbol{X}^T \boldsymbol{\gamma}$,

$$\nabla m(\boldsymbol{Z}, f)[h] = -(Y - \boldsymbol{X}^T \boldsymbol{\beta}) \boldsymbol{X}^T \boldsymbol{\gamma}.$$

Therefore, apply similar decomposition like (S3.5), we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\boldsymbol{X}_i^T\widehat{\boldsymbol{\gamma}}_n-\sqrt{n}P_{\boldsymbol{X},Y}\left\{(Y-\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n)\boldsymbol{X}^T\widehat{\boldsymbol{\gamma}}_n\right\}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)-\sqrt{n}\widehat{\boldsymbol{\gamma}}_n^T\Sigma(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)-\boldsymbol{\gamma}_n^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)\right]\\
+&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\boldsymbol{\gamma}_n^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)-\boldsymbol{\gamma}_n^{*T}\Sigma(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)\right]\\
+&\sqrt{n}(\boldsymbol{\gamma}_n^*-\widehat{\boldsymbol{\gamma}}_n)^T\Sigma(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i
\end{aligned}$$

Since Conditions B.1 - B.3 hold, according to equation (96) of Javanmard et al. (2018), we have

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_n^*)^T\widehat{\Sigma}_n(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)\lesssim\min(s_0,s_1)\frac{\log p_n}{\sqrt{n}}. \tag{S5.9}$$

Additionally, equation (89) of Javanmard et al. (2018) implies that

$$P\left(\sqrt{n}\boldsymbol{e}_1^T(\Sigma_n^{-1}\widehat{\Sigma}_n-I)(\boldsymbol{\beta}_n^*-\widehat{\boldsymbol{\beta}}_n)\geq c_2\sqrt{\frac{s_0}{n}}\log p_n\right)\to 0, \tag{S5.10}$$

as n tends to ∞, where c_2 is a constant. Thus, combing equations (S5.8), (S5.9) and (S5.10) and applying Conditions B.4 and B.9, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\boldsymbol{X}_i^T\widehat{\boldsymbol{\gamma}}_n-\sqrt{n}P_{\boldsymbol{X},Y}^n(Y-\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n)\boldsymbol{X}^T\widehat{\boldsymbol{\gamma}}_n\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i+o_p(1).
\end{aligned}$$

Thus, verifying Condition A.6 is equivalent to proving the equicontinuity of the right-hand side of the above expression. This have been proved in Section S3.

Condition A.7 follows naturally from condition B.6 and B.10. Theorem 1 holds. □

S5.1 Simulation based on oracle h_n^*

In this subsection, we will provide more numerical experiments for OSRE based on oracle h_n^*. Note that $\widehat{h}_n$ is an estimator of h_n^*, while h_n^* would have closed-form when Σ is known. h_n^* in the OSRE of θ should be

$$h_n^*(\boldsymbol{X}) = 2\boldsymbol{\beta}^T \Omega \boldsymbol{X},$$

and h_n^* in the OSRE of β_1 and β_k should be

$$h_n^*(\boldsymbol{X}) = \Omega \boldsymbol{X},$$

where $\Omega = \Sigma^{-1}$. Table 1 shows results for oracle OSRE based on 500 replicates for high-dimensional linear models. The cover probability of the estimators are similar with OSRE based on estimated h_n^*, while the oracle OSRE will have smaller SEs and ESEs than OSRE based on estimated h_n^*. When sample size $n = 200$, SEs for the oracle OSREs are about 70% of the SE fro the OSRE based on estimated h_n^*.

Table 1: Results for oracle OSRE based on 500 replicates for high-dimensional linear models.

n	Method	Parameter	Bias	SE	ESE	CP95	CP90	Bias	SE	ESE	CP95	CP90
					(a)					(b)		
100	OSRE	θ	-0.051	0.412	0.372	0.886	0.828	0.055	0.386	0.402	0.944	0.900
	(oracle)	β_1	-0.010	0.195	0.207	0.954	0.920	0.008	0.202	0.208	0.940	0.902
		β_k	0.005	0.194	0.206	0.968	0.902	0.043	0.223	0.208	0.930	0.866
					(c)					(d)		
	OSRE	θ	0.115	0.273	0.271	0.944	0.886	0.022	0.419	0.399	0.922	0.876
	(oracle)	β_1	-0.005	0.202	0.205	0.936	0.894	-0.005	0.212	0.209	0.936	0.894
		β_k	0.010	0.212	0.206	0.942	0.896	0.004	0.211	0.209	0.938	0.892
					(a)					(b)		
200	OSRE	θ	-0.020	0.276	0.279	0.942	0.884	-0.004	0.255	0.276	0.958	0.918
	(oracle)	β_1	0.013	0.147	0.146	0.942	0.892	-0.010	0.135	0.145	0.960	0.914
		β_k	0.002	0.14	0.146	0.948	0.914	0.018	0.155	0.145	0.912	0.864
					(c)					(d)		
	OSRE	θ	0.041	0.171	0.172	0.952	0.892	-0.008	0.294	0.278	0.926	0.858
	(oracle)	β_1	0.007	0.144	0.144	0.950	0.900	0.006	0.152	0.145	0.944	0.900
		β_k	<0.001	0.141	0.145	0.950	0.906	-0.003	0.154	0.146	0.940	0.878

S6 OSRE for high dimensional generalized linear model

This section will consider statistic inference for a single coefficient in a high-dimensional generalized linear model. Consider n i.i.d samples $(\boldsymbol{X}_i, Y_i)$ with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_n})^T \in \mathbb{R}^{p_n}$ and $Y_i \in \{0, 1\}$, where one of $X's$ is one and the others have mean zero for $k > 1$. Y_i follows a binomial distribution with mean

$$P[Y_i = 1 \mid \boldsymbol{X}_i] = G\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right), \tag{S6.1}$$

where G is the link function, $\boldsymbol{\beta}_n^* = (\beta_{n1}^*, \ldots, \beta_{np_n}^*)^T$ is the vector of parameters. Similarly, let $\mathcal{F}_n = \{f(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\beta_j\}$ be a linear functional class equipped with an inner product defined as former section. Suppose $f_{n0}(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j\beta_{nj}^* \in \mathcal{F}_n$ is unique maximizer of $P\{m(\boldsymbol{X}, Y, f)\}$, where $m(\boldsymbol{X}, Y, f)$ is the log-likelihood function defined as

$$m(\boldsymbol{X}, Y, f) = Y_i \log\left[\frac{G\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)}{1 - G\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)}\right] + \log\left[1 - G\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)\right].$$

Clearly, we have

$$P\left\{\nabla m(\boldsymbol{X}_i, Y_i, f_{n0})[h]\right\} = P\left\{l_x'(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}_{n0})h(\boldsymbol{X}_i)\right\},$$

where

$$l_x'(y, x) = y\frac{G'(x)}{G(x)\left(1 - G(x)\right)} - \frac{G'(x)}{1 - G(x)}$$

and

$$P\left\{\nabla^2 m(\boldsymbol{X}_i, Y_i, f_{n0})[h_1, h_2]\right\} = P\left\{l_{xx}''\left(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}_{n0}\right)h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\right\},$$

where

$$l_{xx}''(y, x) = yH_1(x) - H_2(x),$$

$$H_1(x) = \frac{G''(x)G(x)\left(1 - G(x)\right) - \left(1 - 2G(x)\right)G'^2(x)}{G^2(x)\left(1 - G(x)\right)^2}$$

and

$$H_2(x) = \frac{G''(x)(1-G(x)) + G'^2(x)}{(1-G(x))^2}.$$

Without loss of generality, suppose that we are interested in the first component of $\boldsymbol{\beta}_{n0}$. Then, similar to former section, $\mathfrak{F}_n(f) = \beta^*_{n1}$ which is equivalent to $P[\boldsymbol{e}_1^T \Sigma_n^{-1} \boldsymbol{X} f(\boldsymbol{X})]$. Since m is the log-likelihood function, we can construct h_n^* as

$$h_n^*(\boldsymbol{X}, Y) = -(P\left\{g_{1n}^2(\boldsymbol{X})\right\})^{-1} g_{2n}(\boldsymbol{X}, Y), \tag{S6.2}$$

where

$$g_{1n}(\boldsymbol{X}) = X_1 - \pi(X_1 | X_2, X_3, \ldots, X_{p_n}),$$

$$g_{2n}(\boldsymbol{X}, Y) = l''^{-1}_{xx}(Y, \boldsymbol{X}^T \boldsymbol{\beta}_{n0})(X_1 - \pi(X_1 | X_2, X_3, \ldots, X_{p_n})).$$

To see this,

$$\begin{aligned} P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, X_1]\right\} =& (P g_{1n}^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi[X_1 | X_2, X_3, \ldots, X_{p_n}])X_1] \\ =& (P g_{1n}^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi(X_1 | X_2, X_3, \ldots, X_{p_n}))^2] \\ =& 1 \end{aligned}$$

and for $k > 1$,

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, X_k]\right\} = (P g_{1n}^2(\boldsymbol{X}))^{-1} P[(X_1 - \pi[X_1 | X_2, X_3, \ldots, X_{p_n}])X_k] = 0.$$

Consequently, for any $h_n \in \mathcal{F}_n$ with $h_n(\boldsymbol{x}) = \sum_{j=1}^{p_n} x_j \gamma_j$, we obtain

$$P\left\{\nabla^2 m(\boldsymbol{Z}, f_{n0})[h_n^*, h_n]\right\} = \gamma_1 = \nabla \mathfrak{F}(f_{n0})[h_n].$$

In other words, h_n^* satisfies Condition 3.

Without loss of generality, we suppose the parameter we are interested in is the first coordinate of linear parameter. Suppose the vector parameter $\boldsymbol{\beta}_{n0}$ is sparse, then the initial

estimator $\widehat{\boldsymbol{\beta}}_n$ can be estimated using the Lasso method:

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p_n}} \left\{ \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{X}_i, Y_i, \boldsymbol{\beta}) + 2\lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{S6.3}$$

where $l(\boldsymbol{X}_i, Y_i, \boldsymbol{\beta}) = Y_i \log\left[G\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right) / \left(1 - G\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)\right)\right] + \log\left(1 - G\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)\right)$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n}|\beta_j|$ is the l_1-norm on $\mathbb{R}^{p_n}$ and $\lambda \geq 0$ is a penalty parameter.

Similar to the estimation of h_n^* in (S5.2), we can estimate h_n^* in (S6.2) with Lasso.

$$\begin{aligned} \widehat{g}_{n2}(\boldsymbol{X}) &= l''^{-1}_{xx}\left(Y_i, \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\left(X_1 - \widehat{\pi}(X_1|X_2,\ldots,X_{p_n})\right) \\ &= l''^{-1}_{xx}\left(Y_i, \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\left(X_1 - X_{-1}^T\widehat{\boldsymbol{\eta}}_1^*\right), \end{aligned}$$

where X_{-1} is as defined in former section and $\widehat{\boldsymbol{\eta}}_1^*$ should satisfies

$$\widehat{\boldsymbol{\eta}}_1^* = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2n}\|W_{\widehat{\boldsymbol{\beta}}_n}\boldsymbol{X}_1 - W_{\widehat{\boldsymbol{\beta}}_n}\boldsymbol{X}_{-1}\boldsymbol{\eta}\|_2^2 + \widetilde{\lambda}\|\boldsymbol{\eta}\|_1 \right\}, \tag{S6.4}$$

where $W_{\widehat{\boldsymbol{\beta}}_n}$ is diagonal matrix with (i,j)-th element $U(X_i^T\widehat{\boldsymbol{\beta}}_n)$. Similarly, we estimate $Pg_{n1}^2(X)$ using (S5) by replacing $\boldsymbol{X}$ with $W_{\widehat{\boldsymbol{\beta}}_n}\boldsymbol{X}$ as

$$\widehat{\tau}_1^2 = \|W_{\widehat{\boldsymbol{\beta}}_n}\boldsymbol{X}_1 - W_{\widehat{\boldsymbol{\beta}}_n}\boldsymbol{X}_{-1}^T\widehat{\boldsymbol{\eta}}_1\|_2^2/n + \widetilde{\lambda}\|\widehat{\boldsymbol{\eta}}_1\|_1.$$

Consequently, the estimator for h_n^* is given as

$$\widehat{h}_n(\boldsymbol{X}) = -\widehat{g}_{n2}(\boldsymbol{X})/\widehat{\tau}_1^{*2}.$$

Then, the one-step regularized estimator for β_{n01} is

$$\widetilde{\beta}_{n1} = \widehat{\beta}_{n1} + \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_1^{*-2} l''^{-1}_{xx}\left(Y_i, \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\left(X_{i1} - \sum_{j=2}^{p_n} X_{ij}\widehat{\eta}_{1j}\right) l'_x\left(Y_i, \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right),$$

where $\widehat{\beta}_{n1}$ is the first element of the initial Lasso estimator $\widehat{\boldsymbol{\beta}}_n$. The OSRE is exactly the same as the de-biased Lasso estimator for generalized linear model in Van de Geer et al. (2014).

Obviously, when $G(x) = \exp(x)/(1+\exp(x))$,

$$P\left\{\nabla m(\boldsymbol{X}_i, Y_i, f_{n0})[h]\right\} = P\left\{\left(Y_i - \frac{\exp\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)}{1+\exp\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)}\right) h(\boldsymbol{X}_i)\right\},$$

and

$$P\left\{\nabla^2 m(\boldsymbol{X}_i, Y_i, f_{n0})[h_1, h_2]\right\} = -P\left\{\frac{\exp\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)}{\left[1+\exp\left(\sum_{j=1}^{p_n} X_{ij}\beta_{nj}^*\right)\right]^2} h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\right\}.$$

To simplify the notation, let $U_1(x) = \exp(x)/(1+\exp(x))^2$. Then, we have

$$h_n^*(\boldsymbol{X}) = -(P\left\{g_{1n}^2(\boldsymbol{X})\right\})^{-1} g_{2n}(\boldsymbol{X}), \tag{S6.5}$$

where

$$g_{1n}(\boldsymbol{X}) = X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n}),$$

$$g_{2n}(\boldsymbol{X}) = U_1^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}_{n0})(X_1 - \pi(X_1|X_2, X_3, \ldots, X_{p_n})).$$

Then, the one-step regularized estimator for β_{n01} is

$$\widetilde{\beta}_{n1} = \widehat{\beta}_{n1} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\tau}_1^{*-2} U^{-1}(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n)\left(X_{i1} - \sum_{j=2}^{p_n} X_{ij}\widehat{\eta}_{1j}\right)\left(Y_i - \frac{\exp\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)}{1+\exp\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)}\right).$$

S6.1 Theorem and proof of OSRE for high-dimensional generalized linear models

To state the asymptotic properties for the OSRE, we need the following additional assumptions.

B.11 With probability larger than $1 - p^{-c_1}$,

$$\min\left\{G(\boldsymbol{X}_i^T\boldsymbol{\beta}_{n0}), 1 - G(\boldsymbol{X}_i^T\boldsymbol{\beta}_{n0})\right\} \geq c_2,$$

for $1 \leq i \leq n$ and some small positive constant $c_2 \in (0, 1)$.

Theorem 2. *Suppose that Conditions B.1 - B.4, B.9-B.11 hold true. Furthermore,* $\lambda \asymp \sqrt{\log p_n/n}$ *in* (S6.3) *and* $\widetilde{\lambda} \asymp \sqrt{\log p_n/n}$ *in* (S6.4). *Then*

$$\sqrt{n}(\widetilde{\beta}_{n1} - \beta_{n01}) \xrightarrow{p} N(0, c^2),$$

where $c^2 = P(l''_{xx}(Y, \boldsymbol{X}^T\boldsymbol{\beta}_{n0}))\Omega_{11}$, .

Proof. In order to prove this theorem, we need to verify all the conditions we have listed before. Conditions A.1 - A.3 have been verified in former statements. Thus, the remaining conditions to verify are Conditions A.4 - A.7.

Since $\lambda \asymp \sqrt{\log p_n/n}$ and condition B.1, B.2, B.4, B.11 hold, by applying proposition 1 of Guo et al. (2021), we have

$$P\left(\left\|\widehat{\boldsymbol{\beta}}_n - \beta_n^*\right\|_1 \leq Cs_0(\log p_n/n)^{1/2}\right) \geq 1 - p_n^{-c_1} - \exp(-c_1 n),$$

where $C > 0$ is a positive constant. Combing with condition B.2, $\sqrt{n}d^2_{(n)}(\widehat{f}_n, f_{n0}) = o_p(1)$ which verifies Condition A.4. We now

$$\begin{aligned}
&P\left\{\nabla^2 m(Z, f_{n0})[\widehat{h}_n - h_n^*, \widehat{f}_n - f_{n0}]\right\} \\
=& - P\left\{l''_{xx}\left(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}_n^*\right) l''^{-1}_{xx}\left(Y_i, \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\Sigma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)\right\} \\
=& - P\left\{(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\Sigma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)\right\} + o_p\left((\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\Sigma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)\right),
\end{aligned}$$

where $\boldsymbol{\gamma}_n^* = (1, \boldsymbol{\eta}_{n1}^{*T})^T/\tau_{10}^2$, $\boldsymbol{\eta}_{n1}^*$ can maximize $P\|W_{\boldsymbol{\beta}_{n0}}\boldsymbol{X}_1 - W_{\boldsymbol{\beta}_{n0}}\boldsymbol{X}_{-1}\boldsymbol{\eta}\|_2^2$, $\tau_{10}^2 = P\|W_{\boldsymbol{\beta}_{n0}}\boldsymbol{X}_1 - W_{\boldsymbol{\beta}_{n0}}\boldsymbol{X}_{-1}\boldsymbol{\eta}_{n1}^*\|_2^2$ and $\widehat{\boldsymbol{\gamma}}_n = (1, \widehat{\boldsymbol{\eta}}_1^T)^T/\widehat{\tau}_1^2$. Conditions B.1, B.2, B.9 holds, by applying Theorem 3.2 of Van de Geer et al. (2014) and $\widetilde{\lambda} \asymp \sqrt{\log p_n/n}$ consequent $\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*\|_2 = O_p\left(\sqrt{s_1 \log p_n/n}\right)$.

Thus, we have

$$\sqrt{n}|(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\Sigma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)| = O_p\left(\frac{\sqrt{s_0 s_1}\log p_n}{\sqrt{n}}\right). \tag{S6.6}$$

We use $s_0 = O(n^{\alpha_0}/\log p_n)$ and $s_1 = O(n^{\alpha_1}/\log p_n)$ to obtain that $\sqrt{n}|(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\Sigma(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*)| = o_p(1)$. This verifies Condition A.5.

To verify Condition A.6, recalling that for $h(\boldsymbol{X}) = -\boldsymbol{X}^T\boldsymbol{\gamma}$,

$$\nabla m(\boldsymbol{Z}, f)[h] = -P\left\{l_x'\left(Y_i, \boldsymbol{X}_i^T\boldsymbol{\beta}_n^*\right)\boldsymbol{X}^T\boldsymbol{\gamma}\right\}.$$

Therefore, apply similar decomposition like (S3.5), we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(G(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n) - Y_i\right)\boldsymbol{X}_i^T\widehat{\boldsymbol{\gamma}}_n - \sqrt{n}P_{\boldsymbol{X},Y}\left\{(G(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n) - Y)\boldsymbol{X}^T\widehat{\boldsymbol{\gamma}}_n\right\}\\
=\ &\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(G(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n) - G(\boldsymbol{X}_i^T\boldsymbol{\beta}_n^*)\right)\boldsymbol{X}_i^T\widehat{\boldsymbol{\gamma}}_n - \sqrt{n}P_{\boldsymbol{X},Y}\left\{\left(G(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n) - G(\boldsymbol{X}^T\boldsymbol{\beta}_n^*)\right)\boldsymbol{X}^T\widehat{\boldsymbol{\gamma}}_n\right\}\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i\\
=\ &\frac{1}{\sqrt{n}}\sum_{i=1}^n G'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) - \sqrt{n}\widehat{\boldsymbol{\gamma}}_n^TP(G'(\boldsymbol{X}^T\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i\\
=\ &\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\widehat{\boldsymbol{\gamma}}_n^TG'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\gamma}_n^{*T}G'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\right]\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[G'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{\gamma}_n^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\gamma}_n^{*T}P(G'(\boldsymbol{X}^T\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\right]\\
&+\sqrt{n}(\boldsymbol{\gamma}_n^* - \widehat{\boldsymbol{\gamma}}_n)^TP(G'(\boldsymbol{X}^T\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i,
\end{aligned}$$

where $\varepsilon_i = Y_i - e(\boldsymbol{X}_i^T\boldsymbol{\beta}_n^*)$. As we have illustrated in section S5, conditions B.1 - B.3 lead to equation (S5.9). Then

$$\begin{aligned}
&\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n G'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\right|\\
\le&\frac{\sqrt{n}}{4}\left|\frac{1}{n}\sum_{i=1}^n(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^*)^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\right| \lesssim \min(s_0, s_1)\frac{\log p_n}{\sqrt{n}}.
\end{aligned} \tag{S6.7}$$

Let $\boldsymbol{X}_i^* = \boldsymbol{X}_i\sqrt{G'(\boldsymbol{X}_i\boldsymbol{\beta}^*)}$, $\widehat{\Sigma}_n^* = \frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T}$ and $\Sigma_n^* = P(\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T})$, By applying lemma

6.2 of Javanmard and Montanari (2014a) , we have

$$\left\|\Sigma_n^{*-1}\widehat{\Sigma}_n^* - I\right\|_\infty = O_p\left(\sqrt{\frac{\log p_n}{n}}\right).$$

Then, we have

$$\begin{aligned}&\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[G'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{\gamma}_n^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\gamma}_n^{*T}P(G'(\boldsymbol{X}^T\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n)\right]\right\| \\ \leq&\frac{1}{\sqrt{n}}\left|\boldsymbol{e}_1^T(\Sigma_n^{*-1}\widehat{\Sigma}_n^* - I)\right|\|\boldsymbol{\beta}_n^* - \widehat{\boldsymbol{\beta}}_n\|_1 \lesssim \sqrt{\frac{s_0}{n}}\log p_n,\end{aligned} \tag{S6.8}$$

Thus, combing equations (S6.6), (S6.7) and (S6.8) and applying Conditions B.4 and B.9, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_n\right)\boldsymbol{X}_i^T\widehat{\boldsymbol{\gamma}}_n - \sqrt{n}P_{\boldsymbol{X},Y}^n(Y - \boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_n)\boldsymbol{X}^T\widehat{\boldsymbol{\gamma}}_n \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\boldsymbol{\gamma}}_n^T\boldsymbol{X}_i\varepsilon_i + o_p(1).\end{aligned}$$

Thus, verifying Condition A.6 is equivalent to proving the equicontinuity of the right-hand side of the above expression, which have been proved in Section S3.

Condition A.7 follows naturally from condition B.10 and B.11. Here, we have finished the proof of Theorem 2. □

S6.2 Simulation study with high-dimensional generalized linear models

In this section we will conduct simulation study for high-dimensional generalized linear model. We use similar setting as section 4.1 in the main text . In this setting, we generate $p = 100$ covariates consisting of $K \equiv p/q$ groups, each group with q variables. For q variables in the kth group, denoted by $X_{k1}, ..., X_{kq}$, they are generated as

$$X_{kj} = \frac{(w_{kj} + tu_k)}{1+t}, \quad w_{kj} \sim U(0,1), u_k \in U(0,1).$$

In this way, we generate a sequence of blocked covariates. Y follows binominal distribution with mean $\boldsymbol{X}^T\boldsymbol{\beta}^*$. We set $t = 0.5$ so the correlation between any two X's in the same block

Table 2: Simulation results based on 500 replicates for high-dimensional generalized linear models.

n	Case	method	Bias	SE	ESE	CP95	CP90
100	(a)	PSI	0.691	1.221	0.516	0.884	0.792
		OSRE	-0.055	0.464	0.408	0.932	0.866
	(b)	PSI	0.731	1.242	0.576	0.918	0.842
		OSRE	-0.005	0.456	0.441	0.954	0.910
	(c)	PSI	1.519	2.803	0.963	0.93	0.858
		OSRE	-0.045	0.472	0.459	0.958	0.926
	(d)	PSI	6.477	10.017	27366.613	0.948	0.888
		OSRE	-0.087	0.470	0.469	0.956	0.910
200	(a)	PSI	0.310	0.410	0.268	0.748	0.654
		OSRE	-0.003	0.241	0.203	0.908	0.842
	(b)	PSI	0.317	0.438	0.286	0.746	0.652
		OSRE	-0.004	0.279	0.227	0.910	0.842
	(c)	PSI	0.428	0.572	0.332	0.758	0.634
		OSRE	-0.010	0.288	0.263	0.946	0.894
	(d)	PSI	0.572	0.780	0.383	0.728	0.640
		OSRE	-0.032	0.310	0.286	0.938	0.902

is $\rho = 0.2$, but they are independent if from different blocks. Similar situations as those in section 4.1 are considered as follow

(a) $q = 2$ and $\boldsymbol{\beta}^* = (1, 1, 1, 0, \ldots, 0)^T$.

(b) $q = 4$ and $\boldsymbol{\beta}^* = (1, 1, 1, 0, \ldots, 0)^T$.

(c) $q = 4$ and $\boldsymbol{\beta}^* = (1, 1, 1, 1, 0, \ldots, 0)^T$.

(d) $q = 4$ and $\boldsymbol{\beta}^* = (1, 1, 1, 1, 1, 0, \ldots, 0)^T$.

To illustrate our proposed method, we focus on inference for the coefficient of the first

covariate given by β_1^*. We consider sample size $n = 100$ and 200 replicate each scenario 500 times in the simulation study.

To calculate OSRE, the initial estimate for $\boldsymbol{\beta}$'s is based on the generalized Lasso regression. The tuning parameter is selected by minimizing cross-validation error. The estimate for h_n^* is also obtained from a Lasso regression with cross validation, but the tuning parameter is set to be a factor of the cross-validation optimal parameter. Similar to the choice of high dimensional linear model, we choose a factor of 2^{-6} in the simulation. To examine the inference performance of the proposed method, we report the coverage rate of confidence intervals for OSRE. For comparisons, we also report the coverage rate of the ad-hoc post-selection inference (PSI) by treating selected variables in the Lasso method as the only variable in the logistic regression model.

In Table 2, we report the simulation results of Bias, standard errors (SEs), estimated standard errors (ESEs) the coverage probabilities based on $(1-\alpha)$-confidence intervals, where $\alpha = 0.10$ and 0.05, respectively. CP95 represents coverage rates of 95% confidence intervals and CP90 is the results of 90% confidence interval. We use robust estimator to estimate Bias, SE and ESE. Our proposed method has much smaller bias than PSI. Besides, SE and ESE are inflated in case (d) when sample size $n = 100$ even they are calculated through robust methods. As shown in the table, post-selection inference (PSI) performs very poorly in all of the cases even when sample size equals to 200. Instead, the coverage probabilities of the confidence intervals based on OSRE are reasonably close to the nominal levels no matter $n = 100$ or 200 in all of the cases. This illustrates that out proposed method is also valid for high-dimensional generalized liner model.

S7 Additional numerical example for partial linear model

Section 3.2 in the main text illustrates the example of high-dimensional additive model. It will be also interesting to expand the situation to partial linear model. Since X_1 is a covariate with linear effect, thus we suppose

$$Y_i = \mu + X_{i1}\beta_1^* + \sum_{j=2}^{p_n} f_{nj}^*(X_{ij}) + \varepsilon_i,$$

where μ is a constant and ε_i is the error term with mean zero and finite variance σ^2. Thus, X_1 is the linear part and $(X_2, \ldots, X_p)$ are the non-linear parts of Y.

Obviously, the difference between this semi-parametric model and NAM is that X_1 should be used instead of B-spline bases of X_1. Thus, the estimator are similar as the one in the main text by replacing the B-spline bases with X_1 itself. Table 3 shows the relative bias (Bias), standard errors (SE), estimated standard errors (ESE), the coverage rates of OSREs and PSI based on 500 replicates. CP95 is the coverage rates of 95% confidence interval while CP90 is that of 90% confidence intervals. We can find similar results as the those in non-parametric additive models. Both of the methods have small relative biases. While the ESEs of ad-hoc method(PSI) are smaller than SEs. The coverage probabilities of OSREs are reasonably close to the nominal levels for linear or non-linear parameters when sample size $n = 200$. In the meanwhile, PSI has coverage probabilities lower than $(1 - \alpha)$ for all cases.

S8 Additional results for real data application

We fit both linear model and additive model to this data to test whether any significant linear or nonlinear association exists between any gene and TRIM32. All covariates are standardized by their ranges so the values are between 0 and 1. Linear model fitting is the

Table 3: Simulation results based on 500 replication for high-dimensional partial linear models.

Parameter	n	ρ	Method	Bias	SE	ESE	CP90	CP95
β_1^*	200	0	PSI	-0.010	0.543	0.505	0.842	0.912
			OSRE	-0.005	0.84	0.769	0.868	0.928
		0.2	PSI	-0.002	0.825	0.702	0.870	0.924
			OSRE	-0.036	1.154	1.055	0.886	0.944
	400	0	PSI	-0.002	0.371	0.351	0.892	0.948
			OSRE	-0.004	0.619	0.597	0.908	0.954
		0.2	PSI	-0.005	0.530	0.500	0.878	0.912
			OSRE	-0.003	1.062	0.965	0.902	0.952
$\int f_4^{*2}(x)dx$	200	0	TS	-0.005	1.128	0.873	0.854	0.908
			OSRE	0.036	1.598	1.864	0.926	0.96
		0.2	PSI	0.021	1.126	0.890	0.884	0.928
			OSRE	0.049	1.775	1.938	0.916	0.958
	400	0	PSI	0.006	0.746	0.615	0.886	0.938
			OSRE	0.034	1.159	1.189	0.920	0.964
		0.2	PSI	0.007	0.718	0.632	0.884	0.946
			OSRE	0.011	1.236	1.279	0.936	0.958

same as in the first simulation study where the penalty parameter for Lasso estimation is based on cross validation. OSRE for each regression coefficient in the linear model is then calculated as in the simulation study and its variance is estimated using the proposed method. To fit nonparametric additive model, we use cubic splines with six evenly distributed knots in $[0, 1]$ to estimate each additive components. To test the importance of each covariate, we calculate OSRE for the summary of each functional component as $\int f_k^2(x)dx, k = 1, ..., p$. In the estimation, the tuning parameter is chosen using BIC.

Table 4 summarizes the estimated parameters and their associated p-values, which are computed based on normal distributions. Since the parameter we are interested in is the integral of squared functions, the OSRE for the additive model is about the square of the OSRE for the coefficients in the linear model. This table shows that a total of 25 genes are selected by the linear model but the additive identify 13 important genes. Among those genes, 9 are statistically significant in the linear model if using 0.05 as the significance level but 7 are significant in the additive. Only gene 1367777_at is shown to be significantly associated with TRIM32 for both models.

To further see how these selected genes are associated with TRIM32 in the two model, Figure 1 plots locally weighted scatterplot smoothing estimates for the significant variables claimed in the additive model. The plot indicates that both 1368228_at and 1379971_at reveal nonlinear associations with TRIM32. We also observe a clear linear relationship between TRIM32 and 1367777_at, the only gene that is tested to be significant both models.

Table 4: Parameter inference in the microarray data analysis.

Probe	NAM			Lasso		
	OSRE	Standard Error	p-value	OSRE	Standard Error	p-value
1384035_at	8.34e-06	6.29e-05	0.148			
1368136_at	5.52e-06	1.66e-05	<0.001			
1398370_at	4.04e-05	8.09e-05	<0.001			
1376261_at	4.89e-06	1.23e-04	0.665	1.11e-03	4.25e-02	0.775
1379982_at	9.28e-06	7.24e-05	0.162			
1367777_at	8.45e-05	4.68e-04	0.049	9.63e-03	4.03e-02	0.009
1368228_at	9.13e-06	3.61e-05	0.006			
1380137_at	1.18e-06	1.18e-05	0.274	8.40e-03	3.45e-02	0.008
1384139_at	8.95e-06	7.60e-05	0.199			
1379971_at	1.65e-05	4.29e-05	<0.001			
1388491_at	1.18e-05	2.85e-05	<0.001	2.33e-03	4.55e-02	0.576
1375642_at	8.62e-06	3.96e-05	0.018			
1369414_at	1.88e-05	1.54e-04	0.183			
1372674_at				9.37e-03	4.87e-02	0.036
1370205_at				-2.54e-04	4.04e-02	0.945
1373887_at				1.19e-02	5.72e-02	0.024
1389910_at				4.36e-03	3.93e-02	0.226
1382223_at				1.63e-03	5.99e-02	0.767
1377836_at				5.41e-03	6.03e-02	0.328
1368165_at				1.26e-02	6.72e-02	0.041
1369978_at				4.57e-03	5.60e-02	0.373
1367483_at				8.34e-04	4.28e-02	0.832
1372248_at				1.02e-03	4.32e-02	0.797
1373117_at				4.85e-03	3.92e-02	0.177
1372925_at				8.09e-03	4.68e-02	0.059
1390411_at				-6.47e-03	3.33e-02	0.034
1375833_at				-8.18e-03	3.80e-02	0.019
1372443_at				1.83e-03	8.39e-02	0.812
1393736_at				-1.93e-03	5.53e-02	0.704
1378125_at				-1.10e-02	3.58e-02	0.001
1370261_at				2.71e-03	3.15e-02	0.348
1371551_at				-7.62e-03	3.62e-02	0.022
1371752_at				1.34e-03	4.15e-02	0.725

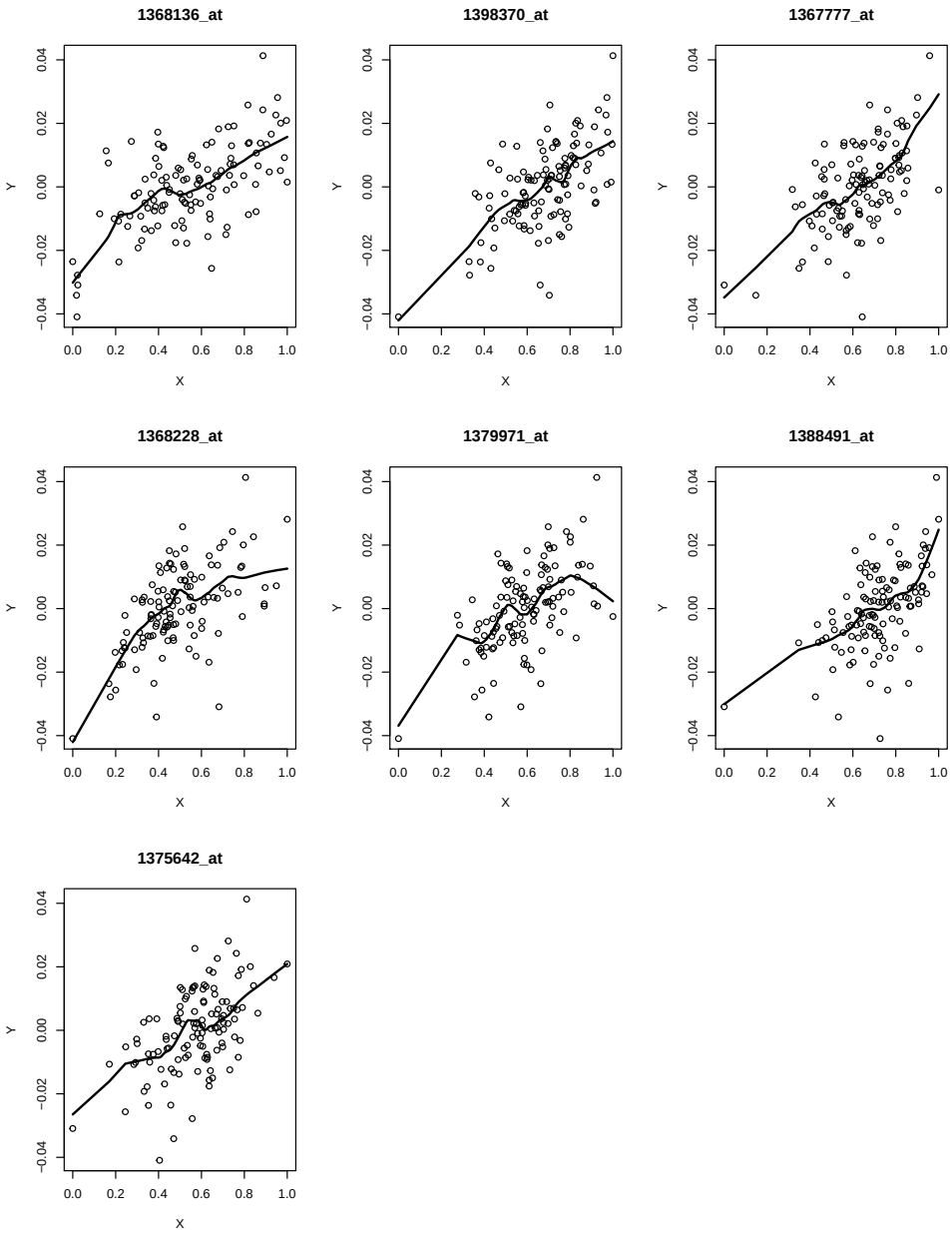

Figure 1: Scatter plots between NAM selected genes and TRIM32.

Bibliography

Bartlett, P. L., S. Mendelson, and J. Neeman (2012). l_1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields 154*(1-2), 193–224.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov

(1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Springer.

De Boor, C., C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor (1978). *A practical guide to splines*, Volume 27. springer-verlag New York.

Guo, Z., P. Rakshit, D. S. Herman, and J. Chen (2021). Inference for the case probability in high-dimensional logistic regression. *The Journal of Machine Learning Research 22*(1), 11480–11533.

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of statistics 38*(4), 2282.

Javanmard, A. and A. Montanari (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Javanmard, A. and A. Montanari (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory 60*(10), 6522–6554.

Javanmard, A., A. Montanari, et al. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer.