Supplementary Material: Sparse Hilbert–Schmidt Independence Criterion Regression

A Asymptotics for U-statistics

U-statistics aim at estimating a parameter, say θ, based on the sample $X_1, \cdots, X_n$. Let us denote by $\psi(.)$ a symmetric kernel that will enter in the U-statistic, which is defined as follows.

Definition. Given a real-valued measurable kernel $\psi(.)$, which is of degreee r and symmetric in its argument, for a sample $X_1, \cdots, X_n$ of size $n \geq r$, then a U-statistic with kernel $\psi(.)$ is defined as

$$U_n = \frac{1}{\binom{n}{r}} \sum_{C_{n,r}} \psi(X_{i_1}, \cdots, X_{i_r}), \tag{6}$$

where the summation is over the set $C_{n,r}$ of all $\binom{n}{r}$ combinations of r integers $i_1 < \cdots < i_r$ chosen from $\{1, \cdots, n\}$.

Note that this definition concerns the particular case of symmetric kernel. Should the kernel be non-symmetric, then the U-statistic would be defined as

$$U_n = \frac{1}{(n)_r} \sum_{i_r^n} \psi(X_{i_1}, \cdots, X_{i_r}),$$

where $(n)_r = \frac{n!}{(n-r)!}$ is the Pochhammer symbol and i_r^n is the index set, which is the set of all r-tuples drawn without replacement from $\{1, \cdots, n\}$.

When deriving the asymptotic distribution of U-statistic based estimator, the variance and covariance of U-statistics are key quantities entering the asymptotic variance covariance matrix. For a given parameter θ and corresponding symmetric kernel $\psi(x_1, \cdots, x_n)$, we consider the class of distribution $\mathcal{F}$ so that $\mathrm{Var}(\psi(X_1, \cdots, X_n)) < \infty$. We then define

$$\psi_r(x_1, \cdots, x_r) = \mathbb{E}_{X_{r+1}, \cdots, X_n}[\psi(x_1, \cdots, x_r, X_{r+1}, \cdots, X_n)],$$

where $X_{r+1}, \cdots, X_n$ are i.i.d. random variables. Then $\psi_0 = \theta$, $\psi_n(x_1, \cdots, x_n) = \psi(x_1, \cdots, x_n)$ and

$$\forall r, \mathbb{E}[\psi_r(X_1, \cdots, X_r)] = \theta.$$

Then to define the variance of the U-statistic, denote by $\sigma_r^2 = \mathrm{Var}(\psi_r(X_1, \cdots, X_r))$.

Theorem A.1. Variance of a U-statistic, (Serfling, 1980)
The variance of a U-statistic given by Eq. (6) *is*

$$\mathrm{Var}(U_n) = \binom{n}{r}^{-1} \sum_{s=1}^{r} \binom{r}{s} \binom{n-r}{r-s} \sigma_s^2. \tag{7}$$

Theorem A.2. Asymptotic distribution of a U-statistic, (Serfling, 1980)
If $\mathrm{Var}(\psi(X_1, \cdots, X_n)) < \infty$ *and* $\sigma_1^2 > 0$, *then*

$$\sqrt{n}\big(U_n - \theta\big) \xrightarrow[n \to \infty]{d} \mathcal{N}_{\mathbb{R}}(0, r^2 \sigma_1^2)$$

Remark. The asymptotic distribution depends on whether $\sigma_1^2 > 0$, in which case we obtain a normal distribution, or $\sigma_1^2 = 0$, in which case the asymptotic law is an infinite linear combination of $\chi^2(1)$ distributions.

Theorem A.3. Covariance of U-statistics, (Serfling, 1980)
Let $U_n^{(k)}$ and $U_n^{(l)}$ be two U-statistics based on the same sample $\mathbf{X} = (X_1, \cdots, X_n)$, respectively equipped with kernel ψ_k and ψ_l of degree r_k, r_l, respectively ($r_k \leq r_l$). Then the covariance is

$$cov(U_n^{(k)}, U_n^{(l)}) = \binom{n}{r_k}^{-1} \sum_{s=1}^{r_k} \binom{r_l}{s} \binom{n-r_l}{r_k-s} \sigma_s^2, \tag{8}$$

where

$$\begin{aligned} \sigma_s^2 &= cov(\mathbb{E}_{X_{s+1},\cdots,X_{r_k}}[\psi_k(x_1, \cdots, x_s, X_{s+1}, \cdots, X_{r_k})], \\ &\quad \mathbb{E}_{X_{s+1},\cdots,X_{r_l}}[\psi_l(x_1, \cdots, x_s, X_{s+1}, \cdots, X_{r_l})]). \end{aligned}$$

Theorem A.4. Multivariate asymptotic distribution, (Hoeffding, 1948)
Let $(U_n^{(k)})$, $k = 1, \cdots, d$ be a sequence of U-statistics, with mean $U^{(k)}$ and kernel $\psi^{(k)}$ of degreee r_k. Let $\mathbf{U}_n = (U_n^{(1)}, \cdots, U_n^{(d)})^\top$ and $\mathbf{U} = (U^{(1)}, \cdots, U^{(d)})^\top$ both d-dimensional vectors. If $\forall k \leq d, \mathrm{Var}(\psi^{(k)}(X_1, \cdots, X_{r_k})) < \infty$, then

$$\sqrt{n}\big(\mathbf{U}_n - \mathbf{U}\big) \xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbb{R}^d}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\Sigma_{kk}, k \leq d$ (resp. $\Sigma_{kl}, k \leq l \leq d$) is given by Eq. (7) (resp. Eq. (8)).

Theorem A.5. Uniform Law of Large Numbers, (Yeo and Johnson, 2001)
Let $\Theta \subset \mathbb{R}^d$ a compact set and consider the kernel

$$U_n(\theta) = \frac{1}{\binom{n}{r}} \sum_{C_{n,r}} \psi(X_{i_1}, \cdots, X_{i_r}; \boldsymbol{\theta}),$$

where $\theta \in \Theta$. Let

$$1 \leq j \leq r, \psi_j(x_1, \cdots, x_j; \boldsymbol{\theta}) = \mathbb{E}_{X_{j+1},\cdots,X_k}[\psi(x_1, \cdots, x_j, X_{j+1}, \cdots, X_k; \boldsymbol{\theta})],$$

and assume

(i) there is an integrable and symmetric kernel $g(.)$ such that $\forall \theta \in \Theta$ and $\forall \boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_r) \in \mathbb{R}^{r\times p}$, $|\psi(\boldsymbol{x}; \boldsymbol{\theta})| \leq g(\boldsymbol{x})$,

(ii) there is a sequence S_M^r of measurable sets such that $\mathbb{P}(\mathbb{R}^{r\times p} - \bigcup_{r=1}^{\infty} S_M^r) = 0$,

(iii) for each M and $\forall j \leq r$, $\psi_j(x_1, \cdots, x_j; \boldsymbol{\theta})$ is equicontinuous in θ for $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j) \in S_M^j$, where $S_M^r = S_M^j \times S_M^{r-j}$,

then

$$\sup_{\theta\in\Theta} |U_n(\theta) - \mathbb{E}[\psi(\boldsymbol{X}; \boldsymbol{\theta})]| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

In this Supplementary material, we report some technical results we relied on to derive the asymptotic distributions

B Large sample distribution

We used the *convexity argument* to derive the asymptotic distribution in the Lasso case. (Chernozhukov and Hong, 2004), (Chernozhukov, 2005) use this convexity argument to obtain the asymptotic distribution of quantile regression type estimators. This argument relies on the convexity Lemma, which is a key result to obtain an asymptotic distribution when the objective function is not differentiable. It only requires the lower-semicontinuity and convexity of the empirical criterion. The convexity Lemma, as in (Chernozhukov, 2005), proof of Theorem 4.1, can be stated as follows.

Lemma B.1. Convexity Lemma, (Chernozhukov, 2005)
Suppose
(i) a sequence of convex lower-semicontinuous $\mathbb{F}_T : \mathbb{R}^d \to \bar{\mathbb{R}}$ *marginally converges to* $\mathbb{F}_\infty : \mathbb{R}^d \to \bar{\mathbb{R}}$ *over a dense subset of* $\mathbb{R}^d$*;*
(ii) $\mathbb{F}_\infty$ *is finite over a nonempty open set* $E \subset \mathbb{R}^d$*;*
(iii) $\mathbb{F}_\infty$ *is uniquely minimized at a random vector* $\boldsymbol{u}_\infty$*.*
Then

$$\underset{\boldsymbol{z}\in\mathbb{R}^d}{\operatorname{argmin}}\ \mathbb{F}_n(\boldsymbol{z}) \xrightarrow{d} \underset{\boldsymbol{z}\in\mathbb{R}^d}{\operatorname{argmin}}\ \mathbb{F}_\infty(\boldsymbol{z}),\ \textit{that is}\ \boldsymbol{u}_n \xrightarrow{d} \boldsymbol{u}_\infty.$$

As for the SCAD and MCP, due to the non-convexity of the penalty function, we used Lemma 3 of (Umezu *et al.*, 2018), which generalises Lemma 2 of (Hjort and Pollard, 1993) to the case of convex non-penalized loss functions with non-convex penalties. Lemma 2 of (Hjort and Pollard, 1993) allows for deriving consistency and asymptotic normality of estimators that are defined by minimisation of convex criterion functions.

Lemma B.2. (Umezu *et al.*, 2018)
Suppose that $\mathbb{G}_n(\boldsymbol{u})$ *is a strictly convex random function that is approximated by* $\tilde{\mathbb{G}}_n(\boldsymbol{u})$*. Let* $\bar{\boldsymbol{u}}$ *be a subvector of* $\boldsymbol{u}$*, and let* $\zeta(\boldsymbol{u})$ *and* $\eta(\bar{\boldsymbol{u}})$ *be continuous functions such that* $\zeta_n(\boldsymbol{u})$ *and* $\eta_n(\bar{\boldsymbol{u}})$ *converge to* $\zeta(\boldsymbol{u})$ *and* $\eta(\bar{\boldsymbol{u}})$ *uniformly over* $\boldsymbol{u}$ *and* $\bar{\boldsymbol{u}}$ *in any compact set, respectively, and assume that* $\zeta(\boldsymbol{u})$ *is convex and* $\eta(\boldsymbol{0}) = 0$*. In addition, for*

$$\nu_n(\boldsymbol{u}) = \mathbb{G}_n(\boldsymbol{u}) + \zeta_n(\boldsymbol{u}) + \eta_n(\bar{\boldsymbol{u}}),\ \textit{and}\ \tilde{\nu}_n(\boldsymbol{u}) = \tilde{\mathbb{G}}_n(\boldsymbol{u}) + \zeta(\boldsymbol{u}) + \eta(\bar{\boldsymbol{u}}),$$

let $\boldsymbol{u}_n$ *and* $\tilde{\boldsymbol{u}}_n$ *be the* argmin *of* $\nu_n(\boldsymbol{u})$ *and* $\tilde{\nu}_n(\boldsymbol{u})$*, respectively, and assume that* $\tilde{\boldsymbol{u}}_n$ *is unique and* $\tilde{\bar{\boldsymbol{u}}}_n = 0$*. Then, for any* $\epsilon > 0, \delta > 0, \mu > \delta$*, there exists* $\gamma > 0$ *such that*

$$\mathbb{P}(\|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n\| \geq \delta) \leq \mathbb{P}(2\Delta_n(\delta) + \epsilon \geq \Upsilon_n(\delta)) + \mathbb{P}(\|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n\| \geq \mu) + \mathbb{P}(\|\bar{\boldsymbol{u}}_n\| \geq \gamma),$$

where

$$\Delta_n(\delta) = \sup_{\boldsymbol{u}:\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n\|\leq\delta} |\nu_n(\boldsymbol{u}) - \tilde{\nu}_n(\boldsymbol{u})|,\ \Upsilon_n(\delta) = \inf_{\boldsymbol{u}:\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n\|=\delta} |\tilde{\nu}_n(\boldsymbol{u}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n)|.$$

Finally, for the large sample distribution of the Bridge penalized estimator, we relied on Theorem 2.7 of (Kim and Pollard, 1990).

Theorem B.3. (Kim and Pollard, 1990)
Let $\{\mathbb{F}_n\}$ *be a random function into the space of all locally bounded real functions on* $\mathbb{R}^d$*, and* $\boldsymbol{u}_n$ *random mapps into* $\mathbb{R}^d$ *such that*

(i) $\mathbb{F}_n \xrightarrow{d} Q$ *for a Borel measure* Q *concentrated on* $C_{\max}(\mathbb{R}^d)$[1]*;*

(ii) $\boldsymbol{u}_n = O_p(1)$*;*

(iii) $\mathbb{F}_n(\boldsymbol{u}_n) \geq \sup_{\boldsymbol{u}} \{\mathbb{F}_n(\boldsymbol{u})\} - \alpha_n$ *for random variables* (α_n) *of order* $o_p(1)$*.*

Then $\boldsymbol{u}_n \xrightarrow{d} \underset{\boldsymbol{u}}{\operatorname{argmax}}\ \{\mathbb{F}(\boldsymbol{u})\}$ *for a* $\mathbb{F}(\boldsymbol{u})$ *with distribution* Q*.*

C Proofs

Proof of Proposition 4.1

Proof. We first expand the V-statistic based quantity

$$\begin{aligned}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) &= \tfrac{1}{n^2}\sum_{i,j=1}^{n}\left(\left(\boldsymbol{H}_n\boldsymbol{L}\boldsymbol{H}_n\right)_{ij} - \sum_{k=1}^{d}\theta_k\left(\boldsymbol{H}_n\boldsymbol{L}_X^k\boldsymbol{H}_n\right)_{ij}\right)^2 \\ &= \widehat{\mathrm{HSIC}}(Y,Y) - 2\sum_{k=1}^{d}\theta_k\widehat{\mathrm{HSIC}}(X^{(k)},Y) + \sum_{k,l=1}^{d}\theta_k\theta_l\widehat{\mathrm{HSIC}}(X^{(k)},X^{(l)}),\end{aligned}$$

[1] See their page 195 for a definition of this set

where $\widehat{\mathrm{HSIC}}(Y,Y)$ is the empirical estimator of $\mathrm{HSIC}(Y,Y)$ given in Eq. (2). Then using Theorem 3 of (Song *et al.*, 2012), which enables to write the estimator of $\mathrm{HSIC}(Y,X)$ as a U-statistic, we have $\widehat{\mathrm{HSIC}}(Y,Y)$ can be rewritten as

$$\widehat{\mathrm{HSIC}}(Y,Y) = \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} h_1(i,j,q,r), \;\; \text{with} \;\; h_1(i,j,q,r) = \sum_{(s,t,u,v)}^{(i,j,q,r)} L_{st}(L_{st} + L_{uv} - 2L_{su}),$$

where the sum represents all ordered quadruples (s,t,u,v) selected without replacement from (i,j,q,r). In the same manner, denoting $K_{st}^k = \psi(X_s^{(k)}, X_t^{(k)})$, we have for any $k,l \leq d$

$$\begin{aligned}
\widehat{\mathrm{HSIC}}(X^{(k)},Y) &= \tfrac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} h_2(i,j,q,r),\; h_2(i,j,q,r) = \tfrac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} K_{st}^k (L_{st}^Y + L_{uv}^Y - 2L_{su}^Y), \\
\widehat{\mathrm{HSIC}}(X^{(k)},X^{(l)}) &= \tfrac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} h_3(i,j,q,r),\; h_3(i,j,q,r) = \tfrac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} K_{st}^k (K_{st}^l + K_{uv}^l - 2K_{su}^l).
\end{aligned}$$

Thus, plugging these U-statistic based estimators of HSIC, we obtain

$$\begin{aligned}
&\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) \\
&= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} \{h_1(i,j,q,r) - 2\sum_{k=1}^d \theta_k h_2(i,j,q,r) + \sum_{k,l=1}^d \theta_k\theta_l h_3(i,j,q,r)\} \\
&= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} \{L_{st}(L_{st} + L_{uv} - 2L_{su}) - 2\sum_{k=1}^d \theta_k K_{st}^k (L_{st} + L_{uv} - 2L_{su}) \\
&\quad + \sum_{k,l=1}^d \theta_k\theta_l K_{st}^k (K_{st}^l + K_{uv}^l - 2K_{su}^l)\} \\
&= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} \frac{1}{4!} \sum_{(s,t,u,v)}^{(i,j,q,r)} \{\big(L_{st} - \sum_{k=1}^d \theta_k K_{st}^k\big)\big(L_{st} - \sum_{l=1}^d \theta_l K_{st}^k\big) \\
&\quad + \big(L_{st} - \sum_{k=1}^d \theta_k K_{st}^k\big)\big(L_{uv} - \sum_{l=1}^d \theta_l K_{uv}^k\big) + \sum_{k=1}^d \theta_k K_{uv}^k L_{st} - \sum_{k=1}^d \theta_k K_{st}^k L_{uv} \\
&\quad - 2\big[\big(L_{st} - \sum_{k=1}^d \theta_k K_{st}^k\big)\big(L_{su} - \sum_{l=1}^d \theta_l K_{su}^l\big) + \sum_{k=1}^d \theta_k K_{su}^k L_{st} - \sum_{k=1}^d \theta_k K_{st}^k L_{su}\big]\} \\
&:= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \boldsymbol{I}_4^n} \ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}).
\end{aligned}$$

□

Proof of Proposition 4.2

Proof. Let us denote $\boldsymbol{z}_1 = (\boldsymbol{x}_1, \boldsymbol{y}_1)^\top \in \mathbb{R}^{p+q}$. Then we rewrite

$$\begin{aligned}
&\nabla_{\theta_k} \ell(\boldsymbol{z}_1, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0) \\
&= \frac{1}{4!}\big[\sum_{s=1,(t,u,v)}^{(2,3,4)} w_{1,tuv}(\boldsymbol{z}_1;\boldsymbol{\theta}_0) + \sum_{t=1,(s,u,v)}^{(2,3,4)} w_{2,suv}(\boldsymbol{z}_1;\boldsymbol{\theta}_0) + \sum_{u=1,(s,t,v)}^{(2,3,4)} w_{3,stv}(\boldsymbol{z}_1;\boldsymbol{\theta}_0) + \sum_{v=1,(s,t,u)}^{(2,3,4)} w_{4,stu}(\boldsymbol{z}_1;\boldsymbol{\theta}_0)\big] \\
&:= \frac{1}{4!}\big[\Upsilon_1(\boldsymbol{\theta}_0) + \Upsilon_2(\boldsymbol{\theta}_0) + \Upsilon_3(\boldsymbol{\theta}_0) + \Upsilon_4(\boldsymbol{\theta}_0)\big],
\end{aligned}$$

where

$$
\begin{aligned}
& w_{1,tuv}(\boldsymbol{z};\boldsymbol{\theta}_0) \\
&= -2\psi(x_1^{(k)},X_t^{(k)})(\phi(y_1,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_t^{(l)}))-\psi(x_1^{(k)},X_t^{(k)})(\phi(Y_u,Y_v)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_u^{(l)},X_v^{(l)})) \\
&- \psi(X_u^{(k)},X_v^{(k)})(\phi(y_1,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_t^{(l)}))+\psi(X_u^{(k)},X_v^{(k)})\phi(y_1,Y_t)-\psi(x_1^{(k)},X_t^{(k)})\phi(Y_u,Y_v) \\
&+ 2\psi(x_1^{(k)},X_t^{(k)})(\phi(y_1,Y_u)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_u^{(l)}))+2\psi(x_1^{(k)},X_u^{(k)})(\phi(y_1,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_t^{(l)})) \\
&+ \psi(x_1^{(k)},X_u^{(k)})\phi(y_1,Y_t)-\psi(x_1^{(k)},X_t^{(k)})\phi(y_1,Y_u),
\end{aligned}
$$

and

$$
\begin{aligned}
& w_{2,suv}(\boldsymbol{z};\boldsymbol{\theta}_0) \\
&= -2\psi(X_s^{(k)},x_1^{(k)})(\phi(Y_s,y_1)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},x_1^{(l)}))-\psi(X_s^{(k)},x_1^{(k)})(\phi(Y_u,Y_v)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_u^{(l)},X_v^{(l)})) \\
&- \psi(X_u^{(k)},X_v^{(k)})(\phi(Y_s,y_1)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},x_1^{(l)}))+\psi(X_u^{(k)},X_v^{(k)})\phi(Y_s,y_1)-\psi(X_s^{(k)},x_1^{(k)})\phi(Y_u,Y_v) \\
&+ 2\psi(X_s^{(k)},x_1^{(k)})(\phi(Y_s,Y_u)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_u^{(l)}))+2\psi(X_s^{(k)},X_u^{(k)})(\phi(Y_s,y_1)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},x_1^{(l)})) \\
&+ \psi(X_s^{(k)},X_u^{(k)})\phi(Y_s,y_1)-\psi(X_s^{(k)},x_1^{(k)})\phi(Y_s,Y_u),
\end{aligned}
$$

and

$$
\begin{aligned}
& w_{3,stv}(\boldsymbol{z};\boldsymbol{\theta}_0) \\
&= -2\psi(X_s^{(k)},X_t^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_t^{(l)}))-\psi(X_s^{(k)},X_t^{(k)})(\phi(y_1,Y_v)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_v^{(l)})) \\
&- \psi(x_1^{(k)},X_v^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_t^{(l)}))+\psi(x_1^{(k)},X_v^{(k)})\phi(Y_s,Y_t)-\psi(X_s^{(k)},X_t^{(k)})\phi(y_1,Y_v) \\
&+ 2\psi(X_s^{(k)},X_t^{(k)})(\phi(Y_s,y_1)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},x_1^{(l)}))+2\psi(X_s^{(k)},x_1^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_t^{(l)})) \\
&+ \psi(X_s^{(k)},x_1^{(k)})\phi(Y_s,Y_t)-\psi(X_s^{(k)},X_t^{(k)})\phi(Y_s,y_1),
\end{aligned}
$$

and finally

$$
\begin{aligned}
& w_{4,stu}(\boldsymbol{z};\boldsymbol{\theta}_0) \\
&= -2\psi(X_s^{(k)},X_t^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_t^{(l)}))-\psi(X_s^{(k)},X_t^{(k)})(\phi(Y_u,y_1)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_u^{(l)},x_1^{(l)})) \\
&- \psi(X_u^{(k)},x_1^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(x_1^{(l)},X_t^{(l)}))+\psi(X_u^{(k)},x_1^{(k)})\phi(Y_s,Y_t)-\psi(X_s^{(k)},X_t^{(k)})\phi(Y_u,y_1) \\
&+ 2\psi(X_s^{(k)},X_t^{(k)})(\phi(Y_s,Y_u)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_u^{(l)}))+2\psi(X_s^{(k)},X_u^{(k)})(\phi(Y_s,Y_t)-\sum_{l=1}^{d}\theta_{0,l}\psi(X_s^{(l)},X_t^{(l)})) \\
&+ \psi(X_s^{(k)},X_u^{(k)})\phi(Y_s,Y_t)-\psi(X_s^{(k)},X_t^{(k)})\phi(Y_s,Y_u).
\end{aligned}
$$

Then taking the expectation with respect to $\boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r$ and using the independence between $\boldsymbol{Y}$ and $\boldsymbol{X}$, we have

$$
\begin{aligned}
&\mathbb{E}[\Upsilon_1(\boldsymbol{\theta}_0)] \\
&= -2\mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(y_1, Y_t)] + 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})\psi(x_1^{(l)}, X_t^{(l)})] - \mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_u, Y_v)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})\psi(X_u^{(l)}, X_v^{(l)})] - \mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(y_1, Y_t)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})\psi(x_1^{(l)}, X_t^{(l)})] + \mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(y_1, Y_t)] - \mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})\phi(Y_u, Y_v)] \\
&+ 2\mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(y_1, Y_u)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})\psi(x_1^{(l)}, X_u^{(l)})] \\
&+ 2\mathbb{E}[\psi(x_1^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(y_1, Y_t)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(x_1^{(k)}, X_u^{(k)})\psi(x_1^{(l)}, X_t^{(l)})] \\
&+ \mathbb{E}[\psi(x_1^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(y_1, Y_t)] - \mathbb{E}[\psi(x_1^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(y_1, Y_u)],
\end{aligned}
$$

Proceeding the same way for the other expectations, we obtain

$$
\begin{aligned}
&\mathbb{E}[\Upsilon_2(\boldsymbol{\theta}_0)] \\
&= -2\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] + 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})\psi(X_s^{(l)}, x_1^{(l)})] - \mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_u, Y_v)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})\psi(X_u^{(l)}, X_v^{(l)})] - \mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})\psi(X_s^{(l)}, x_1^{(l)}))] + \mathbb{E}[\psi(X_u^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] - \mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_u, Y_v)] \\
&+ 2\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, Y_u)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})\psi(X_s^{(l)}, X_u^{(l)})] \\
&+ 2\mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})\psi(X_s^{(l)}, x_1^{(l)})] \\
&+ \mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] - \mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})\phi(Y_s, Y_u)],
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[\Upsilon_3(\boldsymbol{\theta}_0)] \\
&= -2\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] + 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] \\
&- \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(y_1, Y_v)] + \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(x_1^{(l)}, X_v^{(l)})] - \mathbb{E}[\psi(x_1^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(x_1^{(k)}, X_v^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] + \mathbb{E}[\psi(x_1^{(k)}, X_v^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] - \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(y_1, Y_v)] \\
&+ 2\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, y_1)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(X_s^{(l)}, x_1^{(l)})] + 2\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] \\
&- 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] + \mathbb{E}[\psi(X_s^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] - \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, y_1)],
\end{aligned}
$$

and finally

$$
\begin{aligned}
&\mathbb{E}[\Upsilon_4(\boldsymbol{\theta}_0)] \\
&= -2\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] + 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] \\
&- \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_u, y_1)] + \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(X_u^{(l)}, x_1^{(l)})] - \mathbb{E}[\psi(X_u^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] \\
&+ \sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_u^{(k)}, x_1^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] + \mathbb{E}[\psi(X_u^{(k)}, x_1^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] - \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_u, y_1)] \\
&+ 2\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, Y_u)] - 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})\psi(X_s^{(l)}, X_u^{(l)})] + 2\mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] \\
&- 2\sum_{l=1}^{d}\theta_{0,l}\mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})\psi(X_s^{(l)}, X_t^{(l)})] + \mathbb{E}[\psi(X_s^{(k)}, X_u^{(k)})]\mathbb{E}[\phi(Y_s, Y_t)] - \mathbb{E}[\psi(X_s^{(k)}, X_t^{(k)})]\mathbb{E}[\phi(Y_s, Y_u)].
\end{aligned}
$$

Thus by symmetry of the kernel, i.i.d. random variables and under $\mathbb{P}_{\boldsymbol{YX}} = \mathbb{P}_{\boldsymbol{X}}\mathbb{P}_{\boldsymbol{Y}}$, then we obtain $\forall \boldsymbol{z}_1 \in \mathbb{R}^{p+q}$ by summing up all these expectations

$$\forall k \leq d, \mathbb{E}_{\boldsymbol{Z}_j\boldsymbol{Z}_q\boldsymbol{Z}_r}[\nabla_{\theta_k}\ell(\boldsymbol{z}_1, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0)] = 0.$$

□

Proof of Theorem 4.3

Proof. In a first step, we prove the uniform convergence of $\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; .)$ to the limit quantity $\mathbb{L}_\infty^{pen}(.)$ on any compact set $\boldsymbol{B} \subset \Theta$, idest

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{B}} |\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) - \mathbb{L}_\infty^{pen}(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{9}$$

where $\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) = \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) + \sum_{k=1}^{d}\varphi(\frac{\lambda_n}{n}, |\theta_k|)$. We define $\mathcal{C} \subset \Theta$ an open convex set and pick $\boldsymbol{\theta} \in \mathcal{C}$. Then we need to apply the uniform law of large numbers on the non-penalized term, where we rely on Theorem 1 of (Yeo and Johnson, 2001), provided in the technical results (see the Supplementary material). To do so, let $k \leq 4$ and define

$$\ell_k(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}_{k+1}, \cdots, \boldsymbol{Z}_4}[\ell(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k, \boldsymbol{Z}_{k+1}, \cdots, \boldsymbol{Z}_4; \boldsymbol{\theta})].$$

We then need to prove

(i) there is an integrable and symmetric kernel $g(.)$ such that $\forall \theta \in \Theta$ and $\forall \boldsymbol{z} \in \mathbb{R}^{(p+q)\times 4}$, $|\ell(\boldsymbol{z}; \boldsymbol{\theta})| \leq g(\boldsymbol{z})$.

(ii) There is a sequence S_M^4 of measurable sets such that $\mathbb{P}(\mathbb{R}^{(p+q)\times 4} - \bigcup_{M=1}^{\infty} S_M^4) = 0$.

(iii) For each M and $\forall k \leq 4$, $\ell_k(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k; \boldsymbol{\theta})$ is equicontinuous in θ for $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k \in S_M^k$, where $S_M^4 = S_M^k \times S_M^{4-k}$

First, $\ell(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_4; \boldsymbol{\theta})$ is a symmetric kernel that is bounded by assumption 4 and continuous for all $\boldsymbol{\theta}, (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_4) \in \Theta \times \mathbb{R}^{4(p+q)}$. Let $S_M^4 = \bigtimes_{k=1}^4 S_{k,M}$, where $S_{k,M} = \{\boldsymbol{z}_k : |\boldsymbol{z}_{k,i}| \leq M\}$, with M a positive integer. Then we have $\mathbb{P}(\mathbb{R}^{4(p+q)} - \bigcup_{M=1}^{\infty} S_M^4) = 0$. For each fixed M, if $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k) \in S_M^k$, then the coefficients of $\ell_k(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k; \boldsymbol{\theta})$ are bounded, hence $\ell_k(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_k; \boldsymbol{\theta})$ is equicontinuous in $\boldsymbol{\theta}$. As a consequence, we obtain by Theorem 1 of (Yeo and Johnson, 2001) the uniform convergence

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{B}} |\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{x}) - \mathbb{L}_\infty(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Thus, using $\lambda_n/n \to \lambda_0$ and since the parameter is taken over a compact set, we obtain

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{B}} |\mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}) - \mathbb{L}_\infty^{pen}(\boldsymbol{\theta})| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Now we would like

$$\text{argmin } \{\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;.) + \sum_{k=1}^{d}\varphi(\frac{\lambda_n}{n},.)\} \xrightarrow[n\to\infty]{\mathbb{P}} \text{argmin } \{\mathbb{L}_\infty(.) + \sum_{k=1}^{d}\varphi(\lambda_0,.)\}.$$

First, $\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}) + \sum_{k=1}^{d}\varphi(\frac{\lambda_n}{n},|\theta_k|) \geq \mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta})$, and $\arg\min_{\boldsymbol{\theta}\in\Omega} \{\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta})\} = O_p(1)$ by convexity of the criterion, it thus follows that $\arg\min_{\boldsymbol{\theta}\in\Omega} \{\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}) + \sum_{k=1}^{d}\varphi(\frac{\lambda_n}{n},|\theta_k|)\} = O_p(1)$ with probability one. □

Proof of Theorem 4.4

Proof. Let $\nu_n = n^{-1/2} + \sqrt{\text{card}(\mathcal{A})}A_{1,n}$. We would like to prove that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\frac{1}{\nu_n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_\epsilon) < \epsilon.$$

Following the reasoning of (Fan and Li, 2001) Theorem 1, and denoting $\mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}) = \mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}) + \sum_{k=1}^{d}\varphi(\frac{\lambda_n}{n},|\theta_k|)$, we have

$$\mathbb{P}(\frac{1}{\nu_n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_\epsilon) \leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0 + \nu_n\boldsymbol{u}) \leq \mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)),$$

which implies that there is a local minimum in the ball $\{\boldsymbol{\theta}_0 + \nu_n\boldsymbol{u}, \|\boldsymbol{u}\|_2 \leq C_\epsilon\}$ so that the minimum $\hat{\boldsymbol{\theta}}$ satisfies $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\nu_n)$. Now by a Taylor expansion of the penalized loss function, we obtain

$$\begin{aligned}\mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0 + \nu_n\boldsymbol{u}) - \mathbb{L}_n^{pen}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) &= \nu_n\boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \tfrac{\nu_n^2}{2}\boldsymbol{u}^\top\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)\boldsymbol{u} \\ &+ \sum_{k=1}^{d}\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k} + \nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\},\end{aligned}$$

since the third derivative vanishes. We want to prove

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon \;:\; &\boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \tfrac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{H}\boldsymbol{u} + \tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0) \\ + \;&\nu_n^{-1}\big(\sum_{k=1}^{d}\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k} + \nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}\big) \leq 0) < \epsilon,\end{aligned} \tag{10}$$

where $\mathcal{R}_n(\boldsymbol{\theta}_0) = \boldsymbol{u}^\top\{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) - \mathbb{H}\}\boldsymbol{u}$. First, element-by-element $\forall k \leq d$, the score is

$$\begin{aligned}\nabla_{\theta_k}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) &= (n)_4^{-1}\sum_{(i,j,q,r)\in\boldsymbol{I}_4^n}\nabla_\theta\ell(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_q,\boldsymbol{Z}_r;\boldsymbol{\theta}_0) \\ &:= (n)_4^{-1}\sum_{(i,j,q,r)\in\boldsymbol{I}_4^n}\frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)}\big(-2K^k_{st}(L_{st} - \sum_{l=1}^{d}\theta_{0,l}K^l_{st}) - K^k_{st}(L_{uv} - \sum_{l=1}^{d}\theta_{0,l}K^l_{uv}) \\ &- K^k_{uv}(L_{st} - \sum_{l=1}^{d}\theta_{0,l}K^l_{st}) + K^k_{uv}L_{st} - K^k_{st}L_{uv} + 2K^k_{st}(L_{su} - \sum_{l=1}^{d}\theta_{0,l}K^l_{su}) + 2K^k_{su}(L_{st} - \sum_{l=1}^{d}\theta_{0,l}K^l_{st}) \\ &+ K^k_{su}L_{st} - K^k_{st}L_{su}\big).\end{aligned}$$

By the Central Limit Theorem 8, since the score based symmetric kernel $\nabla_\theta\ell(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_q,\boldsymbol{Z}_r;\boldsymbol{\theta}_0)$ is of degree 4 and is non-degenerate by assumption , we have

$$\boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) = O_p(n^{-1/2}\boldsymbol{u}^\top\mathbb{M}\boldsymbol{u}),$$

with $\mathbb{M} = \mathbb{E}[\nabla_\theta \ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\nabla_{\theta^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]$ assumed well defined by assumption 5. Moreover, the Hessian is formed with the non-redundant $d(d+1)/2$ elements $1 \leq k, l \leq d$

$$\nabla^2_{\theta_k\theta_l}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) = (n)_4^{-1}\sum_{(i,j,q,r)\in \boldsymbol{I}_4^n}\frac{1}{24}\sum_{(s,t,u,v)}^{(i,j,q,r)}\big(2K^k_{st}K^l_{st} + K^k_{st}K^l_{uv} + K^k_{uv}K^l_{st} - 2K^k_{st}K^l_{su} - 2K^k_{su}K^l_{st}\big).$$

Thus by the law of large numbers for U-statistics (see e.g. Theorem 3 of (Lee, 1990), subsection 3.4.2)

$$\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{H},$$

with $\mathbb{H} = \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]$ so that $\mathcal{R}_n(\boldsymbol{\theta}_0) = o_p(1)$.

We now focus on the penalty terms. First, we show that the penalty functions satisfy assumption 5. The Lasso satisfies

$$\forall k \in \mathcal{A}, \nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = \frac{\lambda_n}{n}\mathrm{sgn}(\theta_{0,k}),\ \nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = 0.$$

For $0 < q < 1$, the Bridge satisfies

$$\begin{aligned}\forall k \in \mathcal{A}, \qquad &\nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|) = \tfrac{\lambda_n}{n}q|\theta_{0,k}|^{q-1}\mathrm{sgn}(\theta_{0,k}),\\ &\nabla^2_{\theta_k\theta_k}\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|) = \tfrac{\lambda_n}{n}q(q-1)|\theta_{0,k}|^{q-2}\mathrm{sgn}(\theta_{0,k}),\end{aligned}$$

thus the second order derivative of the Bridge converges to 0 when $\lambda_n = o(n)$. As for the SCAD, we have

$$\forall k \in \mathcal{A}, \nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = \frac{\lambda_n}{n}\big(\mathbf{1}_{\{|\theta_{0,k}|\leq\frac{\lambda_n}{n}\}} + \frac{(b_{\mathrm{scad}}\frac{\lambda_n}{n} - |\theta_{0,k}|)_+}{(b_{\mathrm{scad}}-1)\frac{\lambda_n}{n}}\mathbf{1}_{\{|\theta_{0,k}|>\frac{\lambda_n}{n}\}}\big).$$

As a consequence, the SCAD penalty is twice continuously differentiable for $\frac{\lambda_n}{n} < |\theta_{0,k}|$, which implies that $\forall k \in \mathcal{A}, \nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = o(1)$. In the MCP case, we have

$$\forall k \in \mathcal{A}, \nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = (\frac{\lambda_n}{n} - \frac{|\theta_{0,k}|}{b_{\mathrm{mcp}}})\mathrm{sgn}(\theta_{0,k})\mathbf{1}_{\{|\theta_{0,k}|\leq b_{\mathrm{mcp}}\frac{\lambda_n}{n}\}}.$$

Under $\lambda_n = o(n)$, we straightforwardly obtain $\nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = o(1)$ for non-zero components. Now for any $k \in \mathcal{A} \subset \{1,\cdots,d\}$, and since the penalties are coordinate-separable, we have

$$\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|) - \varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = \nu_n u_k\mathrm{sgn}(\theta_{0,k})\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) + \frac{\nu_n^2}{2}u_k^2\nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)\big(1+o(1)\big).$$

Hence, using $\varphi(\frac{\lambda_n}{n},0) = 0$, we have

$$|\sum_{k\in\mathcal{A}}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|) - \varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)| \leq \nu_n\|\boldsymbol{u}\|_1 A_{1,n} + \frac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n}\big(1+o(1)\big).$$

Using $\|\boldsymbol{u}\|_1 \leq \sqrt{\mathrm{card}(\mathcal{A})}\|\boldsymbol{u}\|_2$, and under assumption 5, the third derivative being dominated, we obtain

$$|\sum_{k\in\mathcal{A}}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|) - \varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)| \leq \nu_n\sqrt{\mathrm{card}(\mathcal{A})}\|\boldsymbol{u}\|_2 A_{1,n} + \frac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n}.$$

Then, denoting $\delta_n = \lambda_{\min}(\mathbb{H})C_\epsilon^2\nu_n$, and using $\frac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]\boldsymbol{u} \geq \delta_n$, we deduce that Eq. (10) can be bounded as

$$\begin{aligned}&\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \tfrac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]\boldsymbol{u} + \tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)\\ &+\nu_n^{-1}\big(\sum_{k=1}^d\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}\big) \leq 0)\\ &\leq \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}^\top|\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)| > \delta_n/6) + \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2 = C_\epsilon|\tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)| > \delta_n/6)\\ &\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2 = C_\epsilon : |\sum_{k=1}^d\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}| > \nu_n\delta_n/6).\end{aligned}$$

We have for n and C_ϵ sufficiently large enough

$$\begin{aligned}&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\sum_{k=1}^{d}\{\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k} + \nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\}| > \nu_n\delta_n/6)\\ &\leq \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \nu_n\sqrt{\text{card}(\mathcal{A})}\|\boldsymbol{u}\|_2 A_{1,n} + \tfrac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n} > \nu_n\delta_n/6) < \epsilon/3.\end{aligned}$$

Moreover, if $\nu_n = n^{-1/2} + \sqrt{\text{card}(\mathcal{A})}A_{1,n}$, for C_ϵ large enough

$$\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}^\top|\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)| > \delta_n/6) \leq \frac{C_\epsilon^2 C_{st}}{n\delta_n^2} \leq \frac{C_{st}}{C_\epsilon^4} < \epsilon/3,$$

where $C_{st} > 0$ is a generic constant. Finally, using $\mathcal{R}_n(\boldsymbol{\theta}_0) = o_p(1)$, we obtain

$$\begin{aligned}&\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : \boldsymbol{u}^\top|\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)| > \delta_n/6) + \mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon|\tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)| > \delta_n/6)\\ &\mathbb{P}(\exists \boldsymbol{u}, \|\boldsymbol{u}\|_2 = C_\epsilon : |\sum_{k=1}^{d}\{\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k} + \nu_n \boldsymbol{u}_k|) - \varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\}| > \nu_n\delta_n/6)\\ &\leq \tfrac{C_{st}}{C_\epsilon^4} + 2\epsilon/3 \leq \epsilon,\end{aligned}$$

for n and C_ϵ large enough. We deduce $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\nu_n)$. □

Proof of Theorem 4.5

Proof. Let $\boldsymbol{u} \in \mathbb{R}^d$ such that $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}$, and the empirical criterion $\mathbb{F}_n(\boldsymbol{u}) = n\{\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) - \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\}$. Note that $\mathbb{F}_n(\boldsymbol{u})$ is minimized at $\hat{u}_n = n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ because $\hat{\boldsymbol{\theta}}$ minimizes $\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta})$. Thus $\hat{u}_n = \underset{\boldsymbol{u}\in\mathbb{R}^d}{\text{argmin}}\ \{\mathbb{F}_n(\boldsymbol{u})\}$.
We first establish the finite distributional convergence of $\mathbb{F}_n(.)$ to $\mathbb{F}_\infty(.)$. Then we separate the proof depending on the penalty function under consideration. We have the expansion

$$\mathbb{F}_n(\boldsymbol{u}) = \sqrt{n}\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^\top \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\boldsymbol{u} + n\big(\sum_{k=1}^{d}\{\varphi(\frac{\lambda_n}{n}, |\theta_{0,k} + u_k/\sqrt{n}|) - \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)\}\big).$$

By theorem A, section 5.5.1. of (Serfling, 1980) for the Central Limit Theorem and by Theorem 3, section 3.4.2 of (Lee, 1990) for the Law of Large Numbers for U-statistics

$$\sqrt{n}\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbb{R}^d}(\boldsymbol{0}, \mathbb{M}), \quad \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{H}.$$

As for the penalty terms, in the MCP and SCAD cases, we proceed as follows. We have

$$n\big(\sum_{k=1}^{d}\{\varphi(\frac{\lambda_n}{n}, |\theta_{0,k} + u_k/\sqrt{n}|) - \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)\}\big) = \zeta_n(\boldsymbol{u}) + \eta_n(\boldsymbol{u}),$$

where using the coordinate-separability property of the penalties

$$\zeta_n(\boldsymbol{u}) = n\sum_{k\in\mathcal{A}}\{\varphi(\frac{\lambda_n}{n}, |\theta_{0,k} + u_k/\sqrt{n}|) - \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)\}, \quad \eta_n(\boldsymbol{u}) = n\sum_{k\in\mathcal{A}^c}\{\varphi(\frac{\lambda_n}{n}, u_k/\sqrt{n})\}.$$

Then we have by a Taylor expansion for the indices $k \in \mathcal{A}$

$$\zeta_n(\boldsymbol{u}) = \sum_{k\in\mathcal{A}}\sqrt{n}\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)u_k \text{sgn}(\theta_{0,k}) + \nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)u_k^2/2(1 + o(1)).$$

Under assumption 5, we have $A_{2,n} = o(1)$. We then need to treat the first order term. For both SCAD and MCP, since their derivatives respectively vanish outside $[-b_{\text{scad}}\frac{\lambda_n}{n}, b_{\text{scad}}\frac{\lambda_n}{n}]$, $[-b_{\text{mcp}}\frac{\lambda_n}{n}, b_{\text{mcp}}\frac{\lambda_n}{n}]$ we have

$$\begin{aligned}\nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|) &= \nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\mathbf{1}_{\{|\theta_{0,k}| \leq b_{\text{scad}}\frac{\lambda_n}{n}\}},\\ \nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|) &= \nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\mathbf{1}_{\{|\theta_{0,k}| \leq b_{\text{mcp}}\frac{\lambda_n}{n}\}},\end{aligned}$$

which implies for any $\epsilon > 0$ and $i \in \mathcal{A}$ that

$$\begin{aligned}\mathbb{P}(\sqrt{n}\nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\mathbf{1}_{\{|\theta_{0,k}| \leq b_{\text{scad}}\frac{\lambda_n}{n}\}} > \epsilon) &\leq \mathbb{P}(|\theta_{0,k}| \leq b_{\text{scad}}\tfrac{\lambda_n}{n}) \to 0,\\ \mathbb{P}(\sqrt{n}\nabla_{\theta_k}\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\mathbf{1}_{\{|\theta_{0,k}| \leq b_{\text{mcp}}\frac{\lambda_n}{n}\}} > \epsilon) &\leq \mathbb{P}(|\theta_{0,k}| \leq b_{\text{mcp}}\tfrac{\lambda_n}{n}) \to 0.\end{aligned}$$

As a consequence, $\sqrt{n}\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|) = o_p(1)$. This is a direct consequence of the unbiasedness property when regularizing large coefficients. Thus $\zeta_n(\boldsymbol{u}) \to 0$ as $n \to \infty$. As for $k \in \mathcal{A}^c$, we have

$$\eta_n(\boldsymbol{u}) = \sum_{k\in\mathcal{A}^c} \sqrt{n}\big(\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n}, |\theta_k|)\big)_{\theta_k=0}|u_k| + \frac{1}{2}\big(\nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n}, |\theta_k|)\big)_{\theta_k=0}u_k^2(1+o(1)),$$

Based on the assumption that $\lim_{x\to 0^+}\nabla_x\varphi(\frac{\lambda_n}{n}, x) = \frac{\lambda_n}{n}$ and $\lambda_n = O(\sqrt{n})$, we deduce

$$\eta_n(\boldsymbol{u}) \to \lambda_0 \sum_{k\in\mathcal{A}^c}|u_k|.$$

As a consequence, by Lemma B.2, where in the latter we take $\mathbb{G}_n(\boldsymbol{u}) = \tilde{\mathbb{G}}_n(\boldsymbol{u}) = \sqrt{n}\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^\top\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\boldsymbol{u}$, we obtain $\underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_n\} \xrightarrow{d} \underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_\infty\}$.

For the Bridge estimator, we have for any index $i \in \mathcal{A}\cup\mathcal{A}^c$ and under the rate $\lambda_n/n^{q/2} \to \lambda_0$, then

$$\begin{aligned}&n\big(\sum_{k=1}^d\{\varphi(\frac{\lambda_n}{n}, |\theta_{0,k} + u_k/\sqrt{n}|) - \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)\}\big)\\ &= \lambda_n\sum_{k=1}^d\{|\theta_{0,k} + u_k/\sqrt{n}|^q - |\theta_{0,k}|^q\} \to \lambda_0\sum_{k=1}^d|u_k|^q\mathbf{1}_{\theta_{0,k}=0}.\end{aligned}$$

Now we need to prove that $\underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_n\} \xrightarrow{d} \underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_\infty\}$ for any $\boldsymbol{u}$ and n sufficiently large. To do so, we use Theorem B.3 of (Kim and Pollard, 1990) (see Supplementary material B) and show $\underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_n\} = O_p(1)$. We first have the expansion

$$\begin{aligned}\mathbb{F}_n(\boldsymbol{u}) &= n\{\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}) - \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\} + \lambda_n\sum_{k=1}^d\{|\theta_{0,k} + u_k/\sqrt{n}|^q - |\theta_{0,k}|^q\}\\ &\geq n\{\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}) - \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\} - \lambda_n\sum_{k=1}^d|u_k/\sqrt{n}|^q\\ &\geq n\{\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}) - \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\} - \big(\lambda_0 + \epsilon\big)\sum_{k=1}^d|u_k/\sqrt{n}|^q := \tilde{\mathbb{F}}_n(\boldsymbol{u}),\end{aligned}$$

where, following the proof of Theorem 3 of (Knight and Fu, 2000), ϵ is such that $\lambda_n/n^{q/2} \leq \lambda_0 + \epsilon$. Then, expanding $n\{\mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \boldsymbol{u}/\sqrt{n}) - \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\}$ in $\tilde{\mathbb{F}}_n(\boldsymbol{u})$, we have

$$\tilde{\mathbb{F}}_n(\boldsymbol{u}) = \sqrt{n}\nabla_\theta\mathbb{L}_n(\theta)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^\top\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\theta)\boldsymbol{u} - \big(\lambda_0 + \epsilon\big)\sum_{k=1}^d|u_k/\sqrt{n}|^q.$$

The second term, which is quadratic in $\boldsymbol{u}$, dominates the term with $|\boldsymbol{u}_i|^q$, which is that the quadratic term grows faster $|\boldsymbol{u}_i|^q$. Hence $\underset{\boldsymbol{u}}{\text{argmin}}\,\{\tilde{\mathbb{F}}_n\} = O_p(1)$, which in turns implies $\underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_n\} = O_p(1)$. We then obtain for the Bridge

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_n\} \xrightarrow{d} \underset{\boldsymbol{u}}{\text{argmin}}\,\{\mathbb{F}_\infty\}.$$

Finally, for the Lasso estimator, we have

$$\begin{aligned}&n\big(\sum_{k=1}^d\{\varphi(\frac{\lambda_n}{n}, |\theta_{0,k} + u_k/\sqrt{n}|) - \varphi(\frac{\lambda_n}{n}, |\theta_{0,k}|)\}\big)\\ &= \lambda_n\sum_{k=1}^d\{|\theta_{0,k} + u_k/\sqrt{n}| - |\theta_{0,k}|\} \to \lambda_0\sum_{k=1}^d\big(u_k\text{sgn}(\theta_{0,k})\mathbf{1}_{\theta_{0,k}\neq 0} + |u_k|\mathbf{1}_{\theta_{0,k}=0}\big).\end{aligned}$$

We thus proved that $\mathbb{F}_n(\boldsymbol{u}) \xrightarrow{d} \mathbb{F}_\infty(\boldsymbol{u})$, for a fixed $\boldsymbol{u}$. Let us observe that

$$\boldsymbol{u}_n^* = \underset{\boldsymbol{u}}{\operatorname{argmin}} \{\mathbb{F}_n(\boldsymbol{u})\},$$

and $\mathbb{F}_n(.)$ admits as a minimizer $\boldsymbol{u}_n^* = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. As $\mathbb{F}_n$ is convex and $\mathbb{F}_\infty$ is continuous, convex and has a unique minimum by assumption 5, then by the convexity Lemma B.1, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \underset{\boldsymbol{u}}{\operatorname{argmin}} \{\mathbb{F}_n\} \xrightarrow{d} \underset{\boldsymbol{u}}{\operatorname{argmin}} \{\mathbb{F}_\infty\}.$$

□

Proof of Theorem 4.6

Proof. Let us define $\theta = (\theta_{\mathcal{A}}^\top, \theta_{\mathcal{A}^c}^\top)^\top$. To prove the support recovery consistency, we show with probability tending to one when $n \to \infty$, under $\|\boldsymbol{\theta}_{\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}}\| = O_p(n^{-1/2})$ and suitable regularization rates depending on the penalty, that

$$\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_{\mathcal{A}}, \mathbf{0}_{\mathcal{A}^c}) = \min_{\|\boldsymbol{\theta}_{\mathcal{A}^c}\| \leq C n^{-1/2}} \{\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_{\mathcal{A}}, \boldsymbol{\theta}_{\mathcal{A}^c})\}. \tag{11}$$

To prove Eq. (11), for any $\sqrt{n}$-consistent $\boldsymbol{\theta}_{\mathcal{A}}$, we show that over the set $\{k \in \mathcal{A}^c, \theta_k : |\theta_k| \leq n^{-1/2}C\}$ for $C > 0$

$$\begin{array}{lll} \nabla_{\theta_k} \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) > 0 & \text{when} & 0 < \theta_k < n^{-1/2}C, \\ \nabla_{\theta_k} \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) < 0 & \text{when} & -n^{-1/2}C < \theta_k < 0, \end{array} \tag{12}$$

with probability converging to 1. For any index $k \in \mathcal{A}^c$, by a Taylor expansion around the true parameter, we have

$$\begin{aligned} \nabla_{\theta_k} \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) &= \nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) + \nabla_{\theta_k} \varphi(\frac{\lambda_n}{n}, |\theta_k|)\text{sgn}(\theta_k) \\ &= \nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) + \nabla^2_{\theta_k \theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\big(\theta_k - \theta_{0,k}\big) + \nabla_{\theta_k} \varphi(\frac{\lambda_n}{n}, |\theta_k|)\text{sgn}(\theta_k). \end{aligned}$$

By the central limit theorem and the law of large numbers, we have

$$\sqrt{n}\nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) = O_p(1), \;\; \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \ell(\boldsymbol{Z}; \boldsymbol{\theta}_0)].$$

We thus obtain for the SCAD and MCP penalties

$$\begin{aligned} \nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \hat{\boldsymbol{\theta}}) &= O_p(n^{-1/2}) + \nabla_{\theta_k} \varphi(\tfrac{\lambda_n}{n}, |\theta_k|)\text{sgn}(\theta_k) \\ &= \tfrac{\lambda_n}{n}\{\tfrac{n}{\lambda_n} \nabla_{\theta_k} \varphi(\tfrac{\lambda_n}{n}, |\theta_k|)\text{sgn}(\theta_k) + O_p(\tfrac{\sqrt{n}}{\lambda_n})\}. \end{aligned}$$

As a consequence, under the condition $\lim_{n\to\infty} \liminf_{x\to 0^+} \frac{n}{\lambda_n} \nabla_x \varphi(\frac{\lambda_n}{n}, x) > 0$ and if the regularization parameter satisfies $\frac{\lambda_n}{n^{1/2}} \to \infty$, we deduce that the sign of the gradient entirely depends on the sign of θ_k. This this proves Eq. (12).

For the Bridge penalty, following the same reasoning as in the SCAD and MCP, the non-penalized terms are of order $O_p(n^{-1/2})$. As for the penalty, we have

$$\nabla_{\theta_k} \varphi(\frac{\lambda_n}{n}, |\theta_k|) = \frac{\lambda_n}{n} q|\theta_k|^{q-1}\text{sgn}(\theta_k) = \frac{\lambda_n}{n^{(q+1)/2}} q|n^{1/2}\theta_k|^{q-1}\text{sgn}(\theta_k).$$

As a consequence, we obtain

$$\nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) = \tfrac{\lambda_n}{n^{(q+1)/2}}\{q|n^{1/2}\theta_k|^{q-1}\text{sgn}(\theta_k) + O_p(\tfrac{n^{q/2}}{\lambda_n})\}.$$

Thus, under the assumption that $\lambda_n / n^{q/2} \to \infty$, the sign of θ_k also entirely determines the sign of the gradient.

We now turn to the asymptotic distribution. We proved that $\hat{\boldsymbol{\theta}}_{\mathcal{A}^c}$ degenerates at $\mathbf{0}_{\mathcal{A}^c}$ with probability approaching one. Now by a Taylor expansion around $\theta_{0,k}$, for $k \in \mathcal{A}$, we have

$$\begin{aligned}
&\nabla_{\theta_k}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\hat{\boldsymbol{\theta}}) + \nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\hat{\boldsymbol{\theta}}_k|)\text{sgn}(\hat{\theta}_k) \\
&= \nabla_{\theta_k}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \sum_{j\in\mathcal{A}}\nabla^2_{\theta_k\theta_j}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)(\hat{\theta}_j-\theta_{0,j}) + \nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)\text{sgn}(\theta_{0,k}) \\
&+ \nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)(\hat{\theta}_k-\theta_{0,k})(1+o(1)).
\end{aligned}$$

Then inverting this relationship and multiplying by $\sqrt{n}$, we obtain in vector form with respect to the elements in $\mathcal{A}$

$$\begin{aligned}
&\sqrt{n}\nabla_{\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) \\
&= \big(-\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) - \mathbf{S}_{n,\mathcal{A}\mathcal{A}}\big)\sqrt{n}\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)_{\mathcal{A}} + \big(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \mathbf{S}_{n,\mathcal{A}\mathcal{A}}\big)^{-1}\mathbf{b}_{n,\mathcal{A}}\},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{b}_{n,\mathcal{A}} &= \big(\nabla_{\theta_1}\varphi(\tfrac{\lambda_n}{n},|\theta_{0,1}|)\text{sgn}(\theta_{0,1}),\cdots,\nabla_{\theta_{k_0}}\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k_0}|)\text{sgn}(\theta_{0,k_0})\big)^\top, \\
\mathbf{S}_{n,\mathcal{A}\mathcal{A}} &= \text{diag}(\nabla^2_{\theta_k\theta_k}\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|), k=1,\cdots,k_0).
\end{aligned}$$

We thus deduce that by the central limit theorem for U-statistics and the Slutsky theorem

$$\big(-\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)-\mathbf{S}_{n,\mathcal{A}\mathcal{A}}\big)\sqrt{n}\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)_{\mathcal{A}}+\big(\nabla^2_{\mathcal{A}\mathcal{A}}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)+\mathbf{S}_{n,\mathcal{A}\mathcal{A}}\big)^{-1}\mathbf{b}_{n,\mathcal{A}}\} \xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbb{R}^{k_0}}(\mathbf{0},\mathbb{M}),$$

with

$$\mathbb{M} = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)].$$

□

Proof of Theorem 4.7

The true sparse support now depends on the sample size so that the sparsity assumptions is defined as follows:

Assumption 7. *Sparsity assumption:* $|\mathcal{A}_n| = k_{0,n} < d_n$ *with* $\mathcal{A}_n = \{1 \le k \le d_n : \theta_{0,k} \neq 0\}$.

We assume the following conditions on the penalty functions.

Assumption 8. $\varphi(\frac{\lambda_n}{n},|.|)$ *is twice continuously differentiable except at the origin. We define*

$$A_{1,n} = \max_{1\le k\le d_n}|\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k})|, \quad A_{2,n} = \max_{1\le k\le d_n}|\nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k})|,$$

so that $A_{1,n} = O(n^{-1/2})$ *and* $A_{2,n} \to 0$.

All quantities now depend on d_n, hence on n and should be indexed by n. We denote $\mathbb{H}_n := \mathbb{H}_n(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla^2\boldsymbol{\theta}\boldsymbol{\theta}^\top\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]$ and $\mathbb{M}_n := \mathbb{M}_n(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]$. To make the reading easier, we do not index other quantities by n, which will be implicit. We now assume the following conditions on the loss function.

Assumption 9. $\mathbb{H}_n$ *and* $\mathbb{M}_n$ *exist.* $\mathbb{H}_n$ *is positive-definite and there exist* b_1, b_2 *with* $0 < b_1 < b_2 < \infty$ *and* c_1, c_2 *with* $0 < c_1 < c_2 < \infty$ *such that, for all* n,

$$b_1 < \lambda_{\min}(\mathbb{M}_n) < \lambda_{\max}(\mathbb{M}_n) < b_2, \quad c_1 < \lambda_{\min}(\mathbb{H}_n) < \lambda_{\max}(\mathbb{H}_n) < c_2,$$

where $\lambda_{\min}(K)$ *(resp.* $\lambda_{\max}(K)$*) is the minimum (resp. maximum) eigenvalue of any positive-definite square matrix* K.

Assumption 10. $\mathbb{E}[\{\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\}^2] < \infty$ *and* $\mathbb{E}[\{\nabla^2_{\boldsymbol{\theta}_k\boldsymbol{\theta}_l}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)\}^2] < \infty$ *for every* d_n *(and then of* n*).*

Proof. We proceed as in the proof of Theorem 4.4. We denote $\nu_n = \sqrt{d_n}\{n^{-1/2} + A_{1,n}\}$ and we would like to prove that, for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_\epsilon) < \epsilon.$$

Now, following the same reasoning as in the proof of Theorem 4.4, denoting $\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) = \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}) + \sum_{k=1}^{d_n} \varphi(\frac{\lambda_n}{n}, |\theta_k|)$, we have

$$\mathbb{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > C_\epsilon) \leq \mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_n}, \|\boldsymbol{u}\|_2 = C_\epsilon : \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \nu_n \boldsymbol{u}) \leq \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)),$$

which implies that there is a local minimum in the ball $\{\boldsymbol{\theta}_0 + \nu_n \boldsymbol{u}, \|\boldsymbol{u}\|_2 \leq C_\epsilon\}$ so that the minimum satisfies $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p(\nu_n)$. Now by a Taylor expansion of the penalized loss function

$$\begin{aligned}\mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0 + \nu_n \boldsymbol{u}) - \mathbb{L}_n^{pen}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) &= \nu_n \boldsymbol{u}^\top \nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) + \tfrac{\nu_n^2}{2} \boldsymbol{u}^\top \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) \boldsymbol{u} \\ &+ \sum_{k=1}^{d_n} \{\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k} + \nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\},\end{aligned}$$

since the third derivative vanishes. We want to prove

$$\begin{aligned}\mathbb{P}(\exists \boldsymbol{u} \in \mathbb{R}^{d_n}, \|\boldsymbol{u}\|_2 = C_\epsilon \quad : \quad & \boldsymbol{u}^\top \nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) + \tfrac{\nu_n}{2} \boldsymbol{u}^\top \mathbb{H} \boldsymbol{u} + \tfrac{\nu_n}{2} \mathcal{R}_n(\boldsymbol{\theta}_0) \\ + \quad & \nu_n^{-1} \big(\sum_{k=1}^{d_n} \{\varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k} + \nu_n u_k|) - \varphi(\tfrac{\lambda_n}{n}, |\theta_{0,k}|)\} \big) \leq 0) < \epsilon,\end{aligned} \tag{13}$$

where $\mathcal{R}_n(\boldsymbol{\theta}_0) = \boldsymbol{u}^\top \{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) - \mathbb{H}\} \boldsymbol{u}$. First, for $a > 0$ and the Markov inequality, we have for the score term

$$\begin{aligned}&\mathbb{P}(\sup_{\boldsymbol{u}: \|\boldsymbol{u}\|_2 = C_\epsilon} |\boldsymbol{u}^\top \nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)| > a) \\ &\leq \mathbb{P}(\sup_{\boldsymbol{u}: \|\boldsymbol{u}\|_2 = C_\epsilon} \|\boldsymbol{u}\|_2 \|\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\|_2 > a) \\ &\leq \mathbb{P}(\|\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\|_2 > \frac{a}{C_\epsilon}) \\ &\leq \big(\frac{C_\epsilon}{a}\big)^2 \mathbb{E}[\|\nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\|_2^2] \\ &\leq \big(\frac{C_\epsilon}{a}\big)^2 \sum_{k=1}^{d_n} \mathbb{E}[\big(\nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)\big)^2] \\ &= \big(\frac{C_\epsilon}{a}\big)^2 (n)_4^{-2} \sum_{(i,j,q,r) \in I_4^n} \sum_{(i',j',q',r') \in I_4^n} \sum_{k=1}^{d_n} \mathbb{E}[\{\nabla_{\boldsymbol{\theta}_k} \ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}_k} \ell(\boldsymbol{Z}_{i'}, \boldsymbol{Z}_{j'}, \boldsymbol{Z}_{q'}, \boldsymbol{Z}_{r'}; \boldsymbol{\theta}_0)\}^2].\end{aligned}$$

Thus, by assumption 10, we deduce

$$\mathbb{P}(\sup_{\boldsymbol{u}: \|\boldsymbol{u}\|_2 = C_\epsilon} |\boldsymbol{u}^\top \nabla_\theta \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0)| > a) \leq \frac{C_\epsilon^2 d_n}{a^2 n} K_1,$$

where $0 < K_1 < \infty$ is some constant.

We now focus on the Hessian. First, $\mathcal{R}_n(\boldsymbol{\theta}_0)$ can be written as

$$\mathcal{R}_n(\boldsymbol{\theta}_0) = \sum_{k,l=1}^{d_n} \boldsymbol{u}_k \boldsymbol{u}_l \{\nabla^2_{\boldsymbol{\theta}_k \boldsymbol{\theta}_l} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) - \mathbb{H}_n\}.$$

We have $\mathbb{E}[\mathcal{R}_n(\boldsymbol{\theta}_0)] = 0$ and the variance is defined as

$$\mathrm{Var}(\mathcal{R}_n(\boldsymbol{\theta}_0)) = (n)_4^{-2} \sum_{(i,j,q,r) \in I_4^n} \sum_{(i',j',q',r') \in I_4^n} \sum_{k,k',l,l'=1}^{d_n} \boldsymbol{u}_k \boldsymbol{u}_l \boldsymbol{u}_{k'} \boldsymbol{u}_{l'} \mathbb{E}[\zeta_{k,l,(i,j,q,r)} \zeta_{k',l',(i',j',q',r')}],$$

where $\zeta_{k,l,(i,j,q,r)} = \nabla^2_{\boldsymbol{\theta}_k\boldsymbol{\theta}_l}\ell(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_q,\boldsymbol{Z}_r;\boldsymbol{\theta}_0) - \mathbb{E}[\nabla^2_{\boldsymbol{\theta}_k\boldsymbol{\theta}_l}\ell(\boldsymbol{Z}_i,\boldsymbol{Z}_j,\boldsymbol{Z}_q,\boldsymbol{Z}_r;\boldsymbol{\theta}_0)]$. Let $b>0$, we have by the Markov inequality and assumption 10

$$\mathbb{P}(|\mathcal{R}_n(\boldsymbol{\theta}_0)|>b) \leq \frac{1}{b^2}\mathbb{E}[\mathcal{R}_n^2(\boldsymbol{\theta}_0)] \leq \frac{K_2\|\boldsymbol{u}\|_2^4 d_n^2}{n} \leq \frac{C_\epsilon^4 K_2 d_n^2}{n},$$

for some constant $K_2>0$. By assumption 9,

$$\boldsymbol{u}^\top \mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)]\boldsymbol{u} \geq \lambda_{\min}(\mathbb{H}_n)\boldsymbol{u}^\top\boldsymbol{u}.$$

We now turn to the penalty terms. For any $k\in\mathcal{A}_n\subset\{1,\cdots,d_n\}$, and since the penalties are coordinate-separable, we have

$$\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) = \nu_n u_k \mathrm{sgn}(\theta_{0,k})\nabla_{\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|) + \frac{\nu_n^2}{2}u_k^2\nabla^2_{\theta_k\theta_k}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)\big(1+o(1)\big).$$

Hence, using $\varphi(\frac{\lambda_n}{n},0)=0$, we have

$$|\sum_{k\in\mathcal{A}_n}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)| \leq \nu_n\|\boldsymbol{u}\|_1 A_{1,n} + \frac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n}\big(1+o(1)\big).$$

Using $\|\boldsymbol{u}\|_1 \leq \sqrt{\mathrm{card}(\mathcal{A}_n)}\|\boldsymbol{u}\|_2$, and under assumption 8, the third derivative being dominated, we obtain

$$|\sum_{k\in\mathcal{A}}\varphi(\frac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\frac{\lambda_n}{n},|\theta_{0,k}|)| \leq \nu_n\sqrt{\mathrm{card}(\mathcal{A}_n)}\|\boldsymbol{u}\|_2 A_{1,n} + \frac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n}.$$

Then, denoting $\delta_n = \lambda_{\min}(\mathbb{H}_n)C_\epsilon^2\nu_n$, using $\frac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]\boldsymbol{u}\geq\delta_n$, we deduce that Eq. (13) can be bounded as

$$\begin{aligned}
&\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : \boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \tfrac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]\boldsymbol{u} + \tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)\\
&+\nu_n^{-1}\big(\sum_{k=1}^{d_n}\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}\big)\leq 0)\\
&\leq \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : \boldsymbol{u}^\top|\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)|>\delta_n/6) + \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon|\tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)|>\delta_n/6)\\
&\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : |\sum_{k=1}^{d_n}\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}|>\nu_n\delta_n/6).
\end{aligned}$$

We have for n and C_ϵ sufficiently large enough

$$\begin{aligned}
&\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : |\sum_{k=1}^{d_n}\{\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}+\nu_n u_k|)-\varphi(\tfrac{\lambda_n}{n},|\theta_{0,k}|)\}|>\nu_n\delta_n/6)\\
&\leq \mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : \nu_n\sqrt{\mathrm{card}(\mathcal{A}_n)}\|\boldsymbol{u}\|_2 A_{1,n} + \tfrac{\nu_n^2}{2}\|\boldsymbol{u}\|_2^2 A_{2,n}>\nu_n\delta_n/6)<\epsilon/3.
\end{aligned}$$

Moreover, if $\nu_n = n^{-1/2}+\sqrt{\mathrm{card}(\mathcal{A}_n)}A_{1,n}$, for C_ϵ large enough

$$\mathbb{P}(\exists\boldsymbol{u}\in\mathbb{R}^{d_n},\|\boldsymbol{u}\|_2=C_\epsilon : \boldsymbol{u}^\top|\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0)|>\delta_n/6) \leq \frac{C_\epsilon^2 d_n C_{st}}{n\delta_n^2},$$

where $C_{st}>0$ is a generic constant. Using

$$\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon|\tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)|>\delta_n/6) \leq \tfrac{L_{st}\nu_n^2 d_n^2}{n\delta_n^2},$$

where $L_{st}>0$ is a generic constant. Thus for C_ϵ fixed sufficiently large, under the assumption $d_n^2=O(n)$, we have

$$\begin{aligned}
&\mathbb{P}(\exists\boldsymbol{u},\|\boldsymbol{u}\|_2=C_\epsilon : \boldsymbol{u}^\top\nabla_\theta\mathbb{L}_n(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n;\boldsymbol{\theta}_0) + \tfrac{\nu_n}{2}\boldsymbol{u}^\top\mathbb{E}[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}\ell(\boldsymbol{Z};\boldsymbol{\theta}_0)]\boldsymbol{u} + \tfrac{\nu_n}{2}\mathcal{R}_n(\boldsymbol{\theta}_0)\\
&\tfrac{C_\epsilon^2 d_n C_{st}}{n\delta_n^2} + \tfrac{L_{st}\nu_n^2 d_n^2}{n\delta_n^2} + \epsilon/3 < \epsilon.
\end{aligned}$$

We deduce $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(\nu_n)$. □

D Derivations

Derivation of the HSIC Lasso

$$
\begin{aligned}
n^2 J(\boldsymbol{\theta}) &= n^2 \sum_{k=1}^{d} \theta_k \widehat{\mathrm{HSIC}}(X^{(k)}, Y) - \frac{n^2}{2} \sum_{k=1}^{d} \sum_{k'=1}^{d} \theta_k \theta_{k'} \widehat{\mathrm{HSIC}}(X^{(k)}, X^{(k')}) \\
&= \sum_{k=1}^{d} \theta_k \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n \boldsymbol{K}^{(k)}) - \frac{1}{2} \sum_{k=1}^{d} \sum_{k'=1}^{d} \theta_k \theta_{k'} \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n \boldsymbol{K}^{(k')}) \\
&= \sum_{k=1}^{d} \theta_k \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n \boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n) - \frac{1}{2} \sum_{k=1}^{d} \sum_{k'=1}^{d} \theta_k \theta_{k'} \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n \boldsymbol{H}_n \boldsymbol{K}^{(k')} \boldsymbol{H}_n) \\
&\propto -\frac{1}{2} \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n \boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n) + \sum_{k=1}^{d} \theta_k \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n \boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n) - \frac{1}{2} \sum_{k=1}^{d} \sum_{k'=1}^{d} \theta_k \theta_{k'} \mathrm{trace}(\boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n \boldsymbol{H}_n \boldsymbol{K}^{(k')} \boldsymbol{H}_n), \\
&= -\frac{1}{2} \| \mathrm{vec}(\boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n) - \sum_{k=1}^{d} \theta_k \mathrm{vec}(\boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n) \|_2^2, \\
&= -\frac{1}{2} \| \boldsymbol{H}_n \boldsymbol{L} \boldsymbol{H}_n - \sum_{k=1}^{d} \theta_k \boldsymbol{H}_n \boldsymbol{K}^{(k)} \boldsymbol{H}_n \|_F^2
\end{aligned}
$$

where we use $\boldsymbol{H}_n \boldsymbol{H}_n = \boldsymbol{H}_n$, $\mathrm{trace}(\boldsymbol{A}^\top \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^\top \mathrm{vec}(\boldsymbol{B})$, and $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{R}^{n^2}, (\boldsymbol{A} \in \mathbb{R}^{n \times n})$ is the vectorization operator.

Derivative of the U-statistics

Element-by-element $\forall k \leq d$, the score evaluated at the true parameter $\boldsymbol{\theta}_0$ is given by

$$
\begin{aligned}
\nabla_{\theta_k} \mathbb{L}_n(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n; \boldsymbol{\theta}_0) &= (n)_4^{-1} \sum_{(i,j,q,r)} \nabla_{\theta_k} \ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta}_0) \\
&= (n)_4^{-1} \sum_{(i,j,q,r) \in \boldsymbol{I}_4^n} \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} \big(-2K_{st}^k (L_{st} - \sum_{l=1}^{d} \theta_{0,l} K_{st}^l) - K_{st}^k (L_{uv} - \sum_{l=1}^{d} \theta_{0,l} K_{uv}^l) \\
&- K_{uv}^k (L_{st} - \sum_{l=1}^{d} \theta_{0,l} K_{st}^l) + K_{uv}^k L_{st} - K_{st}^k L_{uv} + 2K_{st}^k (L_{su} - \sum_{l=1}^{d} \theta_{0,l} K_{su}^l) + 2K_{su}^k (L_{st} - \sum_{l=1}^{d} \theta_{0,l} K_{st}^l) \\
&+ K_{su}^k L_{st} - K_{st}^k L_{su} \big).
\end{aligned}
$$

Kernel expression

$$
\mathbb{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}_i \boldsymbol{Z}_j \boldsymbol{Z}_q \boldsymbol{Z}_r} [\ell(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_q, \boldsymbol{Z}_r; \boldsymbol{\theta})],
$$

which can be expressed in terms of kernel as

$$
\begin{aligned}
\mathbb{L}(\boldsymbol{\theta}) &= \mathbb{E}_{YY'}[\phi(Y, Y')^2] - 2\mathbb{E}_{YY}[\mathbb{E}_{Y'}[\phi(Y, Y')]^2] + \mathbb{E}_{YY'}[\phi(Y, Y')]^2 \\
&- 2\sum_{k=1}^{d} \theta_k \{ \mathbb{E}_{YY'X^{(k)}X^{(k)'}}[\phi(Y, Y')\psi(X^{(k)}, X^{(k)'})] - 2\mathbb{E}_{YX^{(k)}}[\mathbb{E}_{Y'}[\phi(Y, Y')]\mathbb{E}_{X^{(k)'}}[\psi(X^{(k)}, X^{(k)'})]] \\
&+ \mathbb{E}_{YY'}[\phi(Y, Y')]\mathbb{E}_{X^{(k)}X^{(k)'}}[\psi(X^{(k)}, X^{(k)'})]\} + \sum_{k,l=1}^{d} \theta_k \theta_l \{ \mathbb{E}_{X^{(k)}X^{(k)'}X^{(l)}X^{(l)'}}[\psi(X^{(k)}, X^{(k)'})\psi(X^{(l)}, X^{(l)'})] \\
&- 2\mathbb{E}_{X^{(k)}X^{(l)}}[\mathbb{E}_{X^{(k)'}}[\psi(X^{(k)}, X^{(k)'})]\mathbb{E}_{X^{(l)'}}[\psi(X^{(l)}, X^{(l)'})]] + \mathbb{E}_{X^{(k)}X^{(k)'}}[\psi(X^{(k)}, X^{(k)'})]\mathbb{E}_{X^{(l)}X^{(l)'}}[\psi(X^{(l)}, X^{(l)'})]\}.
\end{aligned} \tag{14}
$$