Doubly robust tests of exposure effects under high-dimensional confounding: supplementary material

Oliver Dukes, Vahe Avagyan and Stijn Vansteelandt

A Appendix A

Define $\mathbb{N}_p = \{1, 2, \dots, p\}$. We use $\mathbb{E}_{P_n}[\cdot]$ for taking expectation w.r.t. the local data generating process (DGP), whereas $\mathbb{E}_n[\cdot]$ refers to sample expectations. Similarly, $\mathbb{P}_{P_n}[\cdot]$ and $\text{var}_{P_n}[\cdot]$ denote probabilities and variances taken w.r.t. the local DGP respectively. We define the ℓ_∞ norm of any matrix A as $\|A\|_\infty = \max_{i,j} |A_{ij}|$. For a vector $\delta \in \mathbb{R}^p$ and indices $T \subset (1, \dots, p)$, let δ_T denote the vector where $\delta_{Tj} = \delta_j$ if $j \in T$ and $\delta_{Tj} = 0$ otherwise. Also, $T^C = (1, \dots, p) \setminus T$. For sequences a_n and b_n , we use $a_n \lesssim b_n$ to denote that $a_n \leq Cb_n$ for some constant C ($a_n \gtrsim b_n$ is similarly defined). We view the gradients $\partial U_i(\hat{\eta})/\partial \eta$ and $\partial U_i(\eta_n)/\partial \eta$ as row vectors.

A.1 Proof of Theorem 1

In order to prove Theorem 1, we will rely on the following assumptions.

Assumption 1. (*Data generating process*) *There exist constants $C_1, C_2, C_3 < \infty$, $c_4, c_5 > 0$ and $4 < r < \infty$ such that:*

- (i) $\mathbb{E}_{P_n}\{|Y - m(L; \beta_n)|^4 | L\} \leq C_1$ w.p. 1.
- (ii) $\mathbb{E}_{P_n}\{|Y - m(L; \beta_n)|^r\} \leq C_2$.
- (iii) $\max_{i \leq n} \|L_i\|_\infty \leq C_3 < \infty$ w.p. 1.
- (iv) $c_4 \leq \mathbb{E}_{P_n}[\{A - \pi(L; \gamma_n)\}^2 | L]$ and $c_5 \leq \mathbb{E}_{P_n}[\{Y - m(L; \beta_n)\}^2 | L]$ w.p. 1.

Remark. Assumption 1(i) allows one to bound the conditional variance of $Y - m(L; \beta_n)$ given L and also implies a bound on the variance given A and L . Assumption 1(ii) places a bound on the higher order moments of $Y - m(L; \beta_n)$, and is required to show uniform consistency of the variance estimator of $U(\hat{\eta})$ and uniform asymptotic normality of the test statistic. We note that Assumptions 1(i)-(ii) allow for non-Gaussianity and heteroscedasticity with respect to the error term $Y - m(L; \beta_n)$. Assumption 1(iii) requires L to be restricted to a bounded set, which is an assumption commonly made in the literature (sometimes as a primitive condition for proving consistency of estimators) (van de Geer et al., 2014; Farrell, 2015; Ning and Liu, 2017). Assumption 1(iv) places additional bounds on the conditional variance, and implies a type of ‘positivity’ condition such that there must be some variation in A at different levels of L .

Assumption 2. (*Rates for prediction error with unweighted estimators*)

$$(i) \mathbb{E}_n[\{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\}^2] = O_{P_n}(s_\gamma \log(p \vee n)/n).$$

$$(ii) \mathbb{E}_n[\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] = O_{P_n}(s_\beta \log(p \vee n)/n).$$

Remark. Results 2(i)-2(ii) follow from the results of Belloni and Chernozhukov (2013), Belloni et al. (2014), Farrell (2015) and Belloni et al. (2016) on Lasso and Post-Lasso-based estimators. Rates on quantities like $\mathbb{E}_n[\{\gamma_n^T L_i - \hat{\gamma}^T L_i\}^2]$ also follow from those papers.

Proof. The proof will proceed in four steps. In the first step, we show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i(\hat{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i(\eta_n) + o_{P_n}(1) \tag{A.1}$$

in the second, that

$$\frac{\mathbb{E}_n\{U_i(\eta_n)\}}{\sqrt{\frac{1}{n} \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{A.2}$$

in the third, that

$$\mathbb{E}_n[U_i(\hat{\eta})^2 - \mathbb{E}_n\{U_i(\hat{\eta})\}^2]^{-1} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}^{-1} + o_{P_n}(1) \quad (\text{A.3})$$

Finally, we will use these results to show result (4) in the main paper.

Step 1

Consider the sample mean of $U_i(\hat{\eta})$:

$$\mathbb{E}_n\{U_i(\hat{\eta})\} = \mathbb{E}_n\{U_i(\eta_n) + U_i(\hat{\eta}) - U_i(\eta_n)\}$$

After some algebra, we have

$$\sqrt{n}\mathbb{E}_n\{U_i(\hat{\eta})\} = \sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\} + R_1 + R_2 + R_3$$

where

$$\begin{aligned} R_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}, \\ R_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(L_i; \beta_n)\} \{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\} \\ R_3 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\} \{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\} \end{aligned}$$

We aim to show that R_1 , R_2 and R_3 are all $o_{P_n}(1)$ under model $\mathcal{A} \cap \mathcal{B}$.

For R_1 , under the null as defined in (1),

$$\begin{aligned} &\mathbb{E}_{P_n}\{R_1 | (Y_i, L_i)_{i=1}^n\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [\mathbb{E}_{P_n}\{A_i | (Y_i, L_i)_{i=1}^n\} - \pi(L_i; \gamma_n)] \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\} \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}_{P_n} \{R_1^2 | (Y_i, L_i)_{i=1}^n\} \\
&= \mathbb{E}_n \left(\mathbb{E}_{P_n} [\{A_i - \pi(L_i; \gamma_n)\}^2 | (Y_i, L_i)_{i=1}^n] \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2 \right) \\
&\leq C \mathbb{E}_n [\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2]
\end{aligned}$$

where C is a constant. Furthermore, invoking Assumption 2(ii) and sparsity condition (i), we have

$$C \mathbb{E}_n [\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] = o_{P_n}(1)$$

and $\mathbb{E}_{P_n}[R_1^2] = o(1)$. Hence one can then apply Chebyshev's Inequality to show that $|R_1| = o_{P_n}(1)$.

Similarly, for R_2 ,

$$\begin{aligned}
& \mathbb{E}_{P_n} [R_2^2 | (A_i, L_i)_{i=1}^n] \\
&= \mathbb{E}_n \left[\mathbb{E}_{P_n} [\{Y_i - m(L_i; \beta_n)\}^2 | (A_i, L_i)_{i=1}^n] \{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\}^2 \right] \\
&\leq C \mathbb{E}_n [\{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\}^2],
\end{aligned}$$

where C is a constant. This inequality follows from Assumption 1(i). Invoking Assumption 2(i) and sparsity condition (i), we have

$$C \mathbb{E}_n [\{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\}^2] = o_{P_n}(1)$$

so $\mathbb{E}_{P_n}[R_2^2] = o(1)$ and using Chebyshev's inequality, $|R_2| = o_{P_n}(1)$.

Finally, considering R_3 , by Hölder's inequality

$$\begin{aligned}
& \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\} \{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\} \right| \\
&\leq \sqrt{n} \mathbb{E}_n [\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2]^{1/2} \mathbb{E}_n [\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2]^{1/2}
\end{aligned}$$

Then given the joint sparsity condition (ii) on s_γ and s_β , and Assumptions 2(i) and 2(ii), it follows that

$$\sqrt{n}\mathbb{E}_n[\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2]^{1/2}\mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2]^{1/2} = o_{P_n}(1)$$

Therefore $|R_3| = o_{P_n}(1)$ and we have result (A.1).

Step 2

Under the null, we have that

$$\begin{aligned}\text{var}_{P_n}\{U_i(\eta_n)\} &= \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} \\ &= \mathbb{E}_{P_n}[\{A_i - \pi(L_i; \gamma_n)\}^2\{Y_i - m(L_i; \beta_n)\}^2]\end{aligned}$$

and by Assumptions 1(ii) and 1(iv), $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is bounded away from zero (necessary for the inversion) and above uniformly in n .

For some $\epsilon > 0$, such that $4 + 2\epsilon \leq r$

$$\begin{aligned}\mathbb{E}_{P_n}\{|U_i(\eta_n)|^{2+\epsilon}\} &\leq \mathbb{E}_{P_n}\{|A_i - \pi(L_i; \gamma_n)|^{4+2\epsilon}\}^{1/2}\mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^{4+2\epsilon}\}^{1/2} \\ &\leq C\end{aligned}$$

where C is a constant, by Assumption 1(ii). This verifies the Lyapunov condition, such that using this result (and the fact that $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is finite) one can then invoke the Lyapunov central limit theorem for triangular arrays to get result (A.2). We rely on array asymptotics here in order to allow for the data-generating process to change with n .

Step 3

Since $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is bounded away from zero uniformly in n and $\mathbb{E}_{P_n}\{U_i(\eta_n)\} = 0$, given the previous steps it suffices to show that $\mathbb{E}_n\{U_i(\hat{\eta})^2\} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} + o_{P_n}(1)$. We will

first obtain the result

$$\mathbb{E}_n\{U_i(\eta_n)^2\} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} + o_{P_n}(1) \quad (\text{A.4})$$

We have

$$\begin{aligned} & \mathbb{P}_{P_n} \left[|\mathbb{E}_n\{U_i(\eta_n)^2\} - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}|^2 > \epsilon \right] \\ & \leq \frac{1}{\epsilon^2} \mathbb{E}_{P_n} \left[\left| \frac{1}{n} \sum_{i=1}^n U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} \right|^2 \right] \\ & \leq \frac{1}{\epsilon^2 n^2} \left(2 - \frac{1}{n} \right) \sum_{i=1}^n \mathbb{E}_{P_n} \left[|U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}|^2 \right] \end{aligned}$$

where we first apply Chebyshev's inequality. The second uses the Von Bahr-Esseen inequality: let $q \in [1, 2]$, then for independent mean-zero variables $X_1, \dots, X_n$, we have

$$E \left(\left| \sum_{i=1}^n X_i \right|^q \right) \leq \left(2 - \frac{1}{n} \right) \sum_{i=1}^n E(|X_i|^q)$$

(von Bahr and Esseen, 1965).

Since

$$\begin{aligned} & \mathbb{E}_{P_n} \left([U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}]^2 \right) \\ & = \text{var}_{P_n} \{U_i(\eta_n)^2\} \\ & = \text{var}_{P_n} \left(\{A_i - \pi(L_i; \gamma_n)\}^2 \mathbb{E}_{P_n}[\{Y_i - m(L_i; \beta_n)\}^2 | A_i, L_i] \right) \\ & \quad + \mathbb{E}_{P_n} \left(\{A_i - \pi(L_i; \gamma_n)\}^4 \text{var}_{P_n}[\{Y_i - m(L_i; \beta_n)\}^2 | A_i, L_i] \right) \end{aligned}$$

then firstly

$$\begin{aligned} & \text{var}_{P_n} \left(\{A_i - \pi(L_i; \gamma_n)\}^2 \mathbb{E}_{P_n}[\{Y_i - m(L_i; \beta_n)\}^2 | A_i, L_i] \right) \\ & \leq C \mathbb{E}_{P_n}[\{A_i - \pi(L_i; \gamma_n)\}^2]^2 = O(1) \end{aligned}$$

where C is a constant, using Assumption 1(i). Secondly,

$$\begin{aligned} & \mathbb{E}_{P_n} \left(\{A_i - \pi(L_i; \gamma_n)\}^4 \text{var}_{P_n}[\{Y_i - m(L_i; \beta_n)\}^2 | A_i, L_i] \right) \\ & \leq C \mathbb{E}_{P_n}[\{A_i - \pi(L_i; \gamma_n)\}^4] = O(1) \end{aligned}$$

where C is again a constant, invoking Assumptions 1(i) and 1(ii). Result (A.4) then follows.

It remains to show that

$$\mathbb{E}_n\{U_i(\hat{\eta})^2\} = \mathbb{E}_n\{U_i(\eta_n)^2\} + o_{P_n}(1) \quad (\text{A.5})$$

By adding and subtracting $\mathbb{E}_n[\{A_i - \pi(L_i; \hat{\gamma})\}^2 \{Y_i - m(L_i; \beta_n)\}^2]$ and applying the triangle inequality, then

$$\begin{aligned} & |\mathbb{E}_n[\{A_i - \pi(L_i; \hat{\gamma})\}^2 \{Y_i - m(L_i; \hat{\beta})\}^2 - \{A_i - \pi(L_i; \gamma_n)\}^2 \{Y_i - m(L_i; \beta_n)\}^2]| \\ & \leq |\mathbb{E}_n([\{A_i - \pi(L_i; \hat{\gamma})\}^2 - \{A_i - \pi(L_i; \gamma_n)\}^2] \{Y_i - m(L_i; \beta_n)\}^2)| \\ & \quad + |\mathbb{E}_n([\{Y_i - m(L_i; \hat{\beta})\}^2 - \{Y_i - m(L_i; \beta_n)\}^2] \{A_i - \pi(L_i; \hat{\gamma})\}^2)| \\ & = |R_4| + |R_5| \end{aligned}$$

Looking first at $|R_5|$, after some algebra we have

$$\begin{aligned} & |\mathbb{E}_n([\{Y_i - m(L_i; \hat{\beta})\}^2 - \{Y_i - m(L_i; \beta_n)\}^2] \{A_i - \pi(L_i; \hat{\gamma})\}^2)| \\ & \leq |\mathbb{E}_n[\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2 \{A_i - \pi(L_i; \hat{\gamma})\}^2]| \\ & \quad + |2\mathbb{E}_n[\{Y_i - m(L_i; \beta_n)\} \{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\} \{A_i - \pi(L_i; \hat{\gamma})\}^2]| \\ & = |R_{5a}| + |R_{5b}|. \end{aligned}$$

Then,

$$|R_{5a}| \leq \max_{i \leq n} \{A_i - \pi(L_i; \hat{\gamma})\}^2 \mathbb{E}_n[\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2] = o_{P_n}(1)$$

following Assumption 2(ii), the fact that A is binary and sparsity condition (ii). Furthermore,

$$\begin{aligned} |R_{5b}| & \leq 2 \max_{i \leq n} \{A_i - \pi(L_i; \hat{\gamma})\}^2 \mathbb{E}_n[\{Y_i - m(L_i; \beta_n)\}^2]^{1/2} \\ & \quad \times \mathbb{E}_n[\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2]^{1/2} \end{aligned}$$

As before, one can bound $\max_{i \leq n} \{A_i - \pi(L_i; \hat{\gamma})\}^2$, and $\mathbb{E}_n[\{m(L_i; \hat{\beta}) - m(L_i; \beta_n)\}^2]^{1/2} = o_{P_n}(1)$ by Assumption 2(ii). For $\mathbb{E}_n[\{Y_i - m(L_i; \beta_n)\}^2]^{1/2}$, note that by Assumption 1(ii), $\mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^4\} = O(1)$ and hence $\mathbb{E}_{P_n}[\{Y_i - m(L_i; \beta_n)\}^2]^{1/2} = O(1)$. Then to bound the sample average, by the Von Bahr-Esseen inequality:

$$\begin{aligned} & \mathbb{P}_{P_n} \left(\left| \mathbb{E}_n\{|Y_i - m(L_i; \beta_n)|^4\} - \mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^4\} \right|^q > \epsilon \right) \\ & \leq \frac{1}{\epsilon^q n^q} \left(2 - \frac{1}{n} \right) \sum_{i=1}^n \mathbb{E}_{P_n} \left[\left| |Y_i - m(L_i; \beta_n)|^4 - \mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^4\} \right|^q \right] \end{aligned}$$

for $q \in [1, 2]$. Applying Minkowski's inequality and using Assumption 1(ii):

$$\begin{aligned} & \mathbb{E}_{P_n} \left[\left| |Y_i - m(L_i; \beta_n)|^4 - \mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^4\} \right|^q \right] \\ & \leq \left(\mathbb{E}_{P_n} \{|Y_i - m(L_i; \beta_n)|^{4q}\}^{1/q} + \mathbb{E}_{P_n} [\mathbb{E}_{P_n}\{|Y_i - m(L_i; \beta_n)|^4\}^q]^{1/q} \right)^q \\ & = O(1), \end{aligned}$$

hence $\mathbb{E}_n\{|Y_i - m(L_i; \beta_n)|^4\} = O_{P_n}(1)$ and also $\mathbb{E}_n[\{Y_i - m(L_i; \beta_n)\}^2]^{1/2} = O_{P_n}(1)$. Therefore $|R_5| = o_{P_n}(1)$.

Similarly, for R_4 we have

$$\begin{aligned} |R_4| & \leq \mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2 \{Y_i - m(L_i; \beta_n)\}^2] \\ & \quad + 2 \max_{i \leq n} |A_i - \pi(L_i; \gamma_n)| \mathbb{E}_n\{|Y_i - m(L_i; \beta_n)|^4\}^{1/2} \\ & \quad \times \mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2]^{1/2} \end{aligned}$$

By invoking Assumptions 1(ii), 2(i) and the sparsity condition (i), one can show that the second term on the right hand side of the inequality is $o_{P_n}(1)$. Regarding the first term,

$$\begin{aligned} & \mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2 \{Y_i - m(L_i; \beta_n)\}^2] \\ & \leq \max_{i \leq n} |\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)| \mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_n)\}^2]^{1/2} \\ & \quad \times \mathbb{E}_n\{|Y_i - m(L_i; \beta_n)|^4\}^{1/2} \\ & = o_{P_n}(1) \end{aligned}$$

using Holder's inequality, Assumptions 1(ii), 2(i), the previous result that $\mathbb{E}_n\{|Y_i - m(L_i; \beta_n)|^4\} = O_{P_n}(1)$ and the sparsity condition (i). We have shown (A.5) and result (A.3) follows.

Step 4

Consider a sequence $P_n \in P$ such that for any $t \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} |\mathbb{P}_{P_n}(T_n \leq t) - \Phi(t)| > 0.$$

This directly contradicts the results given above that the test statistic T_n converges to a normal distribution with mean 0 and variance 1 under any subsequence P_n in P . $\square$

A.2 Proof of Theorem 2

In the proofs of Theorems 2 and Theorem 3, we will rely on some additional assumptions. Specifically, we will make use of the following rates:

Assumption 3. (*Rates on error of estimated coefficients*) Recall that $s^* = s_\gamma \vee s_\beta$; then

$$(i) \quad \|\gamma_n - \hat{\gamma}_{BR}\|_1 = O_{P_n}(s^* \sqrt{\log(p \vee n)/n}).$$

$$(ii) \quad \|\beta_n - \hat{\beta}_{BR}\|_1 = O_{P_n}(s^* \sqrt{\log(p \vee n)/n}).$$

$$(iii) \quad \|\gamma_n - \hat{\gamma}_{BR}\|_2 = O_{P_n}(\sqrt{s^* \log(p \vee n)/n}).$$

$$(iv) \quad \|\beta_n - \hat{\beta}_{BR}\|_2 = O_{P_n}(\sqrt{s^* \log(p \vee n)/n}).$$

Assumption 4. (*Rates on prediction error for weighted estimators*)

$$(i) \quad \mathbb{E}_n[\{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma}_{BR})\}^2] = O_{P_n}(s^* \log(p \vee n)/n).$$

$$(ii) \quad \mathbb{E}_n[\{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\}^2] = O_{P_n}(s^* \log(p \vee n)/n).$$

Remark. Our proposed nuisance parameter estimators are obtained via (weighted) ℓ_1 penalized regression. The rates in 3 and 4 again follow from the results of Belloni et al. (2016) on weighted ℓ_1 -penalized regression (e.g. their Theorem 4); see also Ning and Liu (2017).

To obtain these rates in Assumptions 3 and 4, we need assumptions on the order of the penalty (which we exploit in the proof); specifically, we need that

$$\lambda_\gamma = O\left(\sqrt{\frac{\log(p \vee n)}{n}}\right) \quad (\text{A.6})$$

$$\lambda_\beta = O\left(\sqrt{\frac{\log(p \vee n)}{n}}\right). \quad (\text{A.7})$$

These are standard assumption on the order of the penalty level in the literature, when working either under the intersection submodel (Farrell, 2015; Belloni et al., 2016), or the union model (Avagyan and Vansteelandt, 2017).

Proof. Repeating the previous decomposition of

$\sqrt{n}\mathbb{E}_n\{U_i(\hat{\eta}_{BR})\} - \sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\}$, for R_1 , we now have

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \hat{\gamma}_{BR})\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \end{aligned} \quad (\text{A.8})$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\pi(L_i; \hat{\gamma}_{BR}) - \pi(L_i; \gamma_n)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \quad (\text{A.9})$$

Then for (A.8), note that following a Taylor expansion,

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \hat{\gamma}_{BR})\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \\ &= -\sqrt{n}\mathbb{E}_n\left\{\frac{\partial U_i(\hat{\eta}_{BR})}{\partial \beta}\right\}(\beta_n - \hat{\beta}_{BR}) + O_{P_n}(\sqrt{n}\|\beta_n - \hat{\beta}_{BR}\|_2^2) \end{aligned}$$

and by Hölder's inequality,

$$\begin{aligned} & \left| \sqrt{n} \mathbb{E}_n \left\{ \frac{\partial U_i(\hat{\eta}_{BR})}{\partial \beta} \right\} (\beta_n - \hat{\beta}_{BR}) \right| \\ & \leq \sqrt{n} \lambda_\gamma \|\beta_n - \hat{\beta}_{BR}\|_1 \end{aligned}$$

using the Karuhn-Kush-Tucker conditions for the ℓ_1 -penalised estimator of β_0 . Therefore, given (A.6), Assumptions 3(ii), 3(iv) and sparsity condition (iii),

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \hat{\gamma}_{BR})\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \right| = o_{P_n}(1)$$

Considering the other term (A.9), along the same lines as in the proof of Theorem 1, one can show that

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\pi(L_i; \hat{\gamma}_{BR}) - \pi(L_i; \gamma_n)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta}_{BR})\} \right| = o_{P_n}(1)$$

using Hölder's inequality, sparsity condition (iii) and Assumptions 4(i) and 4(ii). Therefore $|R_1| = o_{P_n}(1)$. One can re-apply the argument given immediately above to show that $|R_3| = o_{P_n}(1)$.

By noting that

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - m(L_i; \hat{\beta}_{BR})\} \{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma}_{BR})\} \\ & = -\sqrt{n} \mathbb{E}_n \left\{ \frac{\partial U_i(\hat{\eta}_{BR})}{\partial \gamma} \right\} (\gamma_n - \hat{\gamma}_{BR}) + O_{P_n}(\sqrt{n} \|\gamma_n - \hat{\gamma}_{BR}\|_2^2) \end{aligned}$$

and

$$\begin{aligned} & \left| \sqrt{n} \mathbb{E}_n \left\{ \frac{\partial U_i(\hat{\eta}_{BR})}{\partial \gamma} \right\} (\gamma_n - \hat{\gamma}_{BR}) \right| \\ & \leq \sqrt{n} \lambda_\beta \|\gamma_n - \hat{\gamma}_{BR}\|_1 \end{aligned}$$

one can also repeat the above arguments to show that $|R_2| = o_{P_n}(1)$, given (A.7), Assumptions 3(ii), 3(iv), 4(i), 4(ii) and sparsity condition (iii). Result (A.1) follows immediately and the main result follows by essentially repeating Steps 2-4 from the proof of Theorem 1. $\square$

A.3 Proof of Corollary 1

Proof. As discussed in the main paper, when $E(A|L) = \pi(L; \gamma^*)$ and γ^* is known, the proposal in Section 5 for estimating β reduces to standard (unweighted) PMLE. Then repeating Step 1 of the previous proof,

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma^*)\} \{Y_i - m(L_i; \hat{\beta})\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma^*)\} \{Y_i - m(L_i; \beta_n)\} \\ & \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma^*)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\} \end{aligned}$$

Let us define $R_1^* = \sqrt{n} \mathbb{E}_n[\{A_i - \pi(L_i; \gamma^*)\} \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}]$.

Then,

$$\begin{aligned} & \mathbb{E}_{P_n} \{R_1^{*2} | (Y_i, L_i)_{i=1}^n\} \\ &= \mathbb{E}_n \left(\mathbb{E}_{P_n} [\{A_i - \pi(L_i; \gamma^*)\}^2 | (L_i)_{i=1}^n] \{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2 \right) \\ &\leq C \mathbb{E}_n [\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] \end{aligned}$$

where C is a constant. Invoking 4(ii) and sparsity condition (i), we have

$$C \mathbb{E}_n [\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] = o_{P_n}(1),$$

hence $\mathbb{E}_{P_n}[R_1^{*2}] = o(1)$ and $|R_1^*| = o_{P_n}(1)$ using Chebyshev's Inequality. Note that sparsity condition (iii) has not been invoked. $\square$

A.4 Auxiliary results on weighted estimators

Here, we restrict to settings where $m(L; \beta_n) = \beta_n^T L$ and $\pi(L; \gamma_n) = \text{expit}(\gamma_n^T L)$. In this case, note that $\hat{\gamma}_{BR} = \hat{\gamma}$, since no weights are used in estimating this parameter. In order to make transparent the dependence of the estimator $\hat{\beta}_{BR}$ on the weights, we introduce

the notation $\hat{\beta}(\hat{\gamma})$ for when γ (required for the weights) is estimated from the data and $\hat{\beta}(\gamma_n)$ otherwise. In certain places, for convenience we will use the notation $\hat{\delta} = \hat{\beta}(\hat{\gamma}) - \beta_n$, $\tilde{\delta} = \hat{\beta}(\gamma_n) - \beta_n$, $M = \mathbb{E}_n\{L_i w(L_i; \gamma_n) L_i^T\}$ and $\epsilon = Y - m(L; \beta_n)$.

In what follows, we will give a lemma (key to proving Theorem 3) regarding the quantity $\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})$. This will be helpful for understanding the impact of using estimated weights on the distribution of the test statistic. Several additional assumptions are required:

Assumption 5. (*Fast convergence of estimating equations*)

$$\begin{aligned} & \left\| \mathbb{E}_n[w(L_i; \gamma_n)[Y_i - m\{L_i; \hat{\beta}(\gamma_n)\}]L_i - w(L_i; \hat{\gamma})[Y_i - m\{L_i; \hat{\beta}(\hat{\gamma})\}]L_i \right\|_{\infty} \\ &= O_{P_n}\left(\frac{\sqrt{\log p}}{n}\right) \end{aligned} \quad (\text{A.10})$$

Remark. This assumption requires the difference between the estimating equations for $\hat{\beta}(\hat{\gamma})$ and $\hat{\beta}(\gamma_n)$ to shrink very quickly. In low dimensional settings, this difference is exactly zero by virtue of the estimation procedure. In the high dimensional setting, even stronger results than (A.10) are available for our proposed estimators, if we represent our estimator of β_0 as the solution to estimating equations with a bridge penalty:

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \gamma} U_i(\hat{\eta}_{BR}) + \lambda_{\beta} \delta |\hat{\beta}_{BR}|^{\delta-1} \circ \text{sign}(\hat{\beta}_{BR}).$$

Here, $\circ$ is the Hadamard product operator and $g(a)$ denotes a vector of elements $g(a_j)$, where $g(a_j) = \text{sign}(a_j)$ for $j = 1, \dots, p$ if $a_j \neq 0$ and $g(a_j) \in [-1, 1]$ otherwise. As $\delta \rightarrow 1+$, the penalty term $\lambda_{\gamma} \circ g(\hat{\beta}_{BR})$ corresponds to the subgradient of the ℓ_1 or Lasso norm penalty $\|\gamma\|_1$ with respect to β and hence our procedure amounts to ℓ_1 -penalized

m -estimation. Using this representation of our estimator, we can see that

$$\begin{aligned}
& \mathbb{E}_n[w(L_i; \gamma_n)[Y_i - m\{L_i; \hat{\beta}(\gamma_n)\}]L_i - w(L_i; \hat{\gamma})[Y_i - m\{L_i; \hat{\beta}(\hat{\gamma})\}]L_i] \\
&= \sqrt{n}\lambda_\beta\delta[|\hat{\beta}(\gamma_n)|^{\delta-1} \circ \text{sign}(\hat{\beta}(\gamma_n)) - |\hat{\beta}(\hat{\gamma})|^{\delta-1} \circ \text{sign}(\hat{\beta}(\hat{\gamma}))] \\
&= \sqrt{n}\lambda_\beta\delta(\delta-1)|\hat{\beta}(\hat{\gamma})|^{\delta-2} \circ \text{sign}(\hat{\beta}(\hat{\gamma}))\{\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\} \\
&\quad + O_{P_n}(\sqrt{n}\|\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\|_2^2)
\end{aligned}$$

where the final equality follows from a Taylor expansion around $\hat{\beta}(\hat{\gamma})$. For any finite n , we can choose δ to be close enough to 1 such that

$$\sqrt{n}\lambda_\beta\delta(\delta-1)|\hat{\beta}(\hat{\gamma})|^{\delta-2} \circ \text{sign}(\hat{\beta}(\hat{\gamma}))\{\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\}$$

is negligible, since $\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})$ is assumed not to diverge as $\delta \rightarrow 1+$.

Assumption 6. (*Properties of weights*) For any $t \in \mathbb{R}$ and sequence t_n (for $n = 1, 2, \dots$) satisfying $t_n - t = o(1)$, $|w(t_n) - w(t)| \leq w(t)|t_n - t|$, where $w(t) = \text{expit}(t)\{1 - \text{expit}(t)\}$.

Remark. Ning and Liu (2017) demonstrate that the above holds with weights obtained from a logistic regression model.

Assumption 7. (*High-dimensional model selection*)

Let $\hat{s}_\beta$ denote the cardinality $|\text{support}\{\hat{\beta}(\hat{\gamma})\} \cup \text{support}\{\hat{\beta}(\gamma_n)\}|$; then $\hat{s}_\beta = O_{P_n}(s^*)$.

Remark. This assumption states that the number of non-zero entries common to both $\hat{\beta}(\hat{\gamma})$ and $\hat{\beta}(\gamma_n)$ is of similar order to s^* , and is satisfied when $|\text{support}\{\hat{\beta}(\hat{\gamma})\}| = O_{P_n}(s^*)$; note that it does not require perfect model selection (Belloni et al., 2014).

Before giving Lemma 1, we will also review several regularity conditions necessary for the consistency of ℓ_1 -penalized estimators. This is done for clarity, since they will be invoked in the following proofs. For example, we will use the concentration bound

$$\|\mathbb{E}_n[\{A_i - \pi(L_i; \gamma_n)\}L_i]\|_\infty = O_{P_n}\left(\sqrt{\frac{\log(p \vee n)}{n}}\right); \quad (\text{A.11})$$

this can be shown to hold either using the theory of moderate deviations for self-normalised sums (De la Peña et al., 2009; Belloni et al., 2012), or using sub-Gaussian primitive conditions. Furthermore, restricted eigenvalues conditions are also required. For the estimator $\hat{\beta}(\hat{\gamma})$ we need that for any constant $\zeta \geq 1$, there exists a finite constant $\kappa > 0$ such that

$$\min \left[\frac{\hat{\delta}^T M \hat{\delta}}{\|\hat{\delta}\|_2^2} : \hat{\delta} \in \mathbb{R}^p, \|\hat{\delta}_{TC}\|_1 \leq \zeta \|\hat{\delta}_T\|_1, \hat{\delta} \neq 0 \right] \geq \kappa > 0 \quad (\text{A.12})$$

Restricted eigenvalue conditions are key for deriving the nice properties of many sparse estimators (Bickel et al., 2009).

Remark. If the cone constraint on $\hat{\delta}$ in (A.12) is satisfied (and the equivalent condition is satisfied for $\tilde{\delta}$), then by the triangle inequality,

$$\begin{aligned} \left\| \hat{\beta}(\hat{\gamma})_{TC} - \hat{\beta}(\gamma_n)_{TC} \right\|_1 &= \left\| \hat{\beta}(\hat{\gamma})_{TC} - \beta_{nTC} + \beta_{nTC} - \hat{\beta}(\gamma_n)_{TC} \right\|_1 \\ &\leq \left\| \hat{\delta}_{TC} \right\|_1 + \left\| \tilde{\delta}_{TC} \right\|_1 \\ &\leq \zeta \left(\left\| \hat{\delta}_T \right\|_1 + \left\| \tilde{\delta}_T \right\|_1 \right). \end{aligned}$$

Defining the set

$$\Delta_{\zeta, T} = \left\{ \hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \in \mathbb{R}^p \setminus \{0\} : \left\| \hat{\beta}(\hat{\gamma})_{TC} - \hat{\beta}(\gamma_n)_{TC} \right\|_1 \leq \zeta \left(\left\| \hat{\delta}_T \right\|_1 + \left\| \tilde{\delta}_T \right\|_1 \right) \right\}$$

and the minimum restricted eigenvalue of M as

$$\varphi_{\zeta}^2(M) = \min_{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \in \Delta_{\zeta, T}, |T| \leq s^*} \left[\frac{\{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\}^T M \{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\}}{\left\| \hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \right\|_2^2} \right],$$

condition (A.12) thus implies that

$$\varphi_{\zeta}^2(M) \geq \kappa > 0. \quad (\text{A.13})$$

For a comprehensive discussion of suitable regularity conditions for proving the consistency of ℓ_1 -penalized estimators, we refer the interested reader to Bühlmann and van de Geer (2011).

Lemma 1. *In addition to the sparsity conditions (i) and (iv), suppose that the Assumptions 1(iii), 2(i), 3(i), 4(ii), and 5-7 hold. Then it follows that for a given sequence P_n we have that*

$$\begin{aligned} & \left\| \hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \right\|_1 \\ &= O_{P_n} \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s^* \sqrt{\log(p \vee n)} (s_\gamma \sqrt{\log(p \vee n)} + 1)}{n} \right). \end{aligned} \quad (\text{A.14})$$

Proof. Many of the arguments are similar to those in Appendix E of Ning and Liu (2017).

Let $\bar{\delta} = \hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)$; then

$$\begin{aligned} \bar{\delta}^T M \bar{\delta} &= \mathbb{E}_n \left(w(L_i; \gamma_n) (\bar{\delta}^T L_i) [Y_i - m\{L_i; \hat{\beta}(\gamma_n)\}] \right) \\ &\quad - \mathbb{E}_n \left(w(L_i; \gamma_n) (\bar{\delta}^T L_i) [Y_i - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \right) \\ &= \mathbb{E}_n \left(w(L_i; \gamma_n) (\bar{\delta}^T L_i) [Y_i - m\{L_i; \hat{\beta}(\gamma_n)\}] \right) \\ &\quad - \mathbb{E}_n \left(w(L_i; \hat{\gamma}) (\bar{\delta}^T L_i) [Y_i - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \right) \\ &\quad + \mathbb{E}_n \left[\{w(L_i; \hat{\gamma}) - w(L_i; \gamma_n)\} (\bar{\delta}^T L_i) \{Y_i - m(L_i; \beta_n)\} \right] \\ &\quad + \mathbb{E}_n \left(\{w(L_i; \hat{\gamma}) - w(L_i; \gamma_n)\} (\bar{\delta}^T L_i) [m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \right) \\ &= R_6 + R_7 + R_8 + R_9. \end{aligned}$$

Firstly,

$$\begin{aligned} & |R_6 + R_7| \\ &\leq \|\mathbb{E}_n[w(L_i; \gamma_n)\{Y_i - \tilde{m}(L_i)\}L_i - w(L_i; \hat{\gamma})\{Y_i - \hat{m}(L_i)\}L_i]\|_\infty \|\bar{\delta}\|_1 \\ &\leq C' \frac{\sqrt{\log(p \vee n)}}{n} \|\bar{\delta}\|_1 \end{aligned}$$

where C' is a constant, due to Assumption 5. For R_8 ,

$$\begin{aligned}
|R_8| &= \left| \mathbb{E}_n \left[\left\{ \frac{w(L_i; \hat{\gamma}) - w(L_i; \gamma_n)}{w(L_i; \gamma_n)(\hat{\gamma}^T L_i - \gamma_n^T L_i)} \right\} \sqrt{w(L_i; \gamma_n)} (\bar{\delta}^T L_i) \right. \right. \\
&\quad \left. \left. \times \sqrt{w(L_i; \gamma_n)} (\hat{\gamma}^T L_i - \gamma_n^T L_i) \{Y_i - m(L_i; \beta_n)\} \right] \right| \\
&\leq \max_{i \leq n} |\epsilon_i| (\bar{\delta}^T M \bar{\delta})^{1/2} \mathbb{E}_n [(\hat{\gamma}^T L_i - \gamma_n^T L_i)^2]^{1/2} \\
&\lesssim \max_{i \leq n} |\epsilon_i| (\bar{\delta}^T M \bar{\delta})^{1/2} \sqrt{\frac{s_\gamma \log(p \vee n)}{n}}
\end{aligned}$$

under Assumptions 2 and 6. Next,

$$\begin{aligned}
|R_9| &\leq \mathbb{E}_n [w(L_i; \gamma_n) |\hat{\gamma}^T L_i - \gamma_n^T L_i| |\bar{\delta}^T L_i| |m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\hat{\gamma})\}|] \\
&\leq \max_{i \leq n} \|L_i\|_\infty \|\hat{\gamma} - \gamma_n\|_1 \mathbb{E}_n [w(L_i; \gamma_n) |\bar{\delta}^T L_i| |m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\hat{\gamma})\}|] \\
&\leq \max_{i \leq n} \|L_i\|_\infty \|\hat{\gamma} - \gamma_n\|_1 \\
&\quad \times (\bar{\delta}^T M \bar{\delta})^{1/2} \mathbb{E}_n \left([m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\hat{\gamma})\}]^2 \right)^{1/2} \\
&\lesssim (\bar{\delta}^T M \bar{\delta})^{1/2} \left(\frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right)
\end{aligned}$$

invoking Assumptions 1(iii), 3(i), 4(ii) and 6.

Putting this together,

$$\begin{aligned}
\bar{\delta}^T M \bar{\delta} &\leq C' \frac{\sqrt{\log p}}{n} \|\bar{\delta}\|_1 \\
&\quad + (\bar{\delta}^T M \bar{\delta})^{1/2} C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right) \tag{A.15}
\end{aligned}$$

where C^* is a constant. Let us consider two cases; firstly, assume

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \leq C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right)$$

holds. Note (A.13) implies that

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \geq \varphi_\zeta(M) \|\bar{\delta}\|_2 \gtrsim \frac{1}{\sqrt{\hat{s}_\beta}} \|\bar{\delta}\|_1.$$

Invoking Assumption 7 and combining the lower and upper bound,

$$\|\bar{\delta}\|_1 \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s_\gamma s^* \log(p \vee n)}{n} \quad (\text{A.16})$$

On the other hand, if

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \geq C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^* \log(p \vee n)}}{n} \right)$$

then rearranging (A.15), it follows that

$$\begin{aligned} & (\bar{\delta}^T M \bar{\delta})^{1/2} \left\{ (\bar{\delta}^T M \bar{\delta})^{1/2} - C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^* \log(p \vee n)}}{n} \right) \right\} \\ & \leq C' \frac{\sqrt{\log(p \vee n)}}{n} \|\bar{\delta}\|_1 \end{aligned}$$

Using (A.13), we have

$$\begin{aligned} & (\bar{\delta}^T M \bar{\delta})^{1/2} - C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^* \log(p \vee n)}}{n} \right) \\ & \lesssim \frac{\sqrt{s^* \log(p \vee n)}}{n} \end{aligned}$$

so

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^* \log(p \vee n)} + \sqrt{s^* \log(p \vee n)}}{n}$$

and again by (A.13),

$$\|\bar{\delta}\|_1 \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s^* \sqrt{\log(p \vee n)} (s_\gamma \sqrt{\log(p \vee n)} + 1)}{n}. \quad (\text{A.17})$$

Taking the union of the bounds (A.16) and (A.17) completes the proof. $\square$

A.5 Proof of Theorem 3

Assumption 8. (*Regularity conditions on the errors*)

$\max_{i \leq n} |\epsilon_i| \sqrt{s_\gamma s^* \log(p \vee n)} = o(\sqrt{n})$ w.p. 1.

Remark. This can be most simply shown to hold if $\max_{i \leq n} |\epsilon_i| = O_{P_n}(1)$. Belloni et al. (2012) and Farrell (2015) suppose that $\max_{i \leq n} |\epsilon_i| = O_{P_n}(n^{1/r})$ for some $r > 2$, such that larger values of r allow one to relax assumptions on sparsity in exchange for stronger conditions on the distributions of the errors. If the ϵ_i are normal, then r can be arbitrarily large. Alternatively, one can place the stronger sub-Gaussian conditions on ϵ_i , whereby $\max_{i \leq n} |\epsilon_i| = O_{P_n}(\sqrt{\log n})$.

We now give the proof.

Proof. Decomposing $\sqrt{n}\mathbb{E}_n\{U_i(\hat{\eta}_{BR})\} - \sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\}$ as in the proof of Theorem 1, one can show $|R_2| = o_{P_n}(1)$ along the lines of the proof of Theorem 1, appealing to Assumptions 1(i), 2(i) and sparsity condition (i). Similarly, one can show that R_3 is $o_{P_n}(1)$ using the joint sparsity condition (ii) and Assumptions 2(i) and 4(ii).

Then for R_1 ,

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} [m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} [m(L_i; \beta_n) - m\{L_i; \hat{\beta}(\gamma_n)\}] \\ & \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} [m\{L_i; \hat{\beta}(\gamma_n)\} - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \\ &= R_{1a} + R_{1b} \end{aligned}$$

One can show $|R_{1a}| = o_{P_n}(1)$ using Assumption 4(ii) and sparsity condition (i). For R_{1b} ,

$$\begin{aligned} & \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \{A_i - \pi(L_i; \gamma_n)\} [m\{L_i; \hat{\beta}(\gamma_n)\} - m\{L_i; \hat{\beta}(\hat{\gamma})\}] \right| \\ & \leq \sqrt{n} \|\mathbb{E}_n[\{A_i - \pi(L_i; \gamma_n)\} L_i]\|_\infty \left\| \hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma}) \right\|_1 \\ &= O_{P_n}(\sqrt{\log(p \vee n)}) \\ & \quad \times O_{P_n} \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s^* \sqrt{\log(p \vee n)} (s_\gamma \sqrt{\log(p \vee n)} + 1)}{n} \right) \end{aligned}$$

by (A.11) and Lemma 1 given in Appendix A.4. Hence under sparsity assumptions (i), (iv) and Assumption 8, $|R_{1b}| = o_{P_n}(1)$ and thus $R_1 = o_{P_n}(1)$. By then repeating Steps 2-4 from the proof of Theorem 1, the main result follows. $\square$

Remark. We note that it follows from the above proof that sharper results are also available in linear models under misspecification than are given in Theorem 2. Namely, when either model $\mathcal{A}$ or the linear model for $\mathcal{B}$ is misspecified, ultra-sparsity is only required in the correct model. For example, if $E(A|L) = \pi(L; \gamma)$, then we require $s_\gamma^2 = o(n)$ but only $s_\beta = o(n)$ (ignoring log factors).

B Appendix B

Below, we describe an iterative method for γ and β (based on the reasoning in Section 5), when both are parameters indexing logistic models. In practice, one can take the penalty terms obtained via cross validation during the first iteration of the algorithm ($j = 1$) and use the same terms in subsequent iterations.

C Appendix C

Here we include some additional simulation results. Compared with the setting considered in Section 6 of the main paper, we allowed for a more dense model for the exposure A . Specifically $g = \left(40 \frac{\log(20)}{n^{1/2}}, 40 \frac{\log(19)}{n^{1/2}}, \dots, 40 \frac{\log(2)}{n^{1/2}}, 0_{20}, 0_{21}, 2 \frac{\log(2)}{n^{1/2}}, \dots, 2 \frac{\log(60)}{n^{1/2}}, 0_{81}, \dots, 0_p\right)$ now. Data generation process under the alternative outcome model (to evaluate the impact of fitting a misspecified linear model) was the same as in the main paper.

We further relaxed the assumption in Section 6 of homoscedastic errors. Specifically, the outcome Y_i was normally distributed conditional on L_i with mean $1 + \beta_i L_i$ and standard deviation $\sqrt{\left(\frac{1 + \beta_i L_i}{E_n((1 + \beta_i L_i)^2)}\right)}$. In settings where we fitted a misspecified outcome model

you cannot have a sample average in the data-generating mechanism

Algorithm 1 An algorithm for estimating η when Y is binary

1. Estimate γ and β as $\hat{\gamma}^{(0)}$ and $\hat{\beta}^{(0)}$ using (unweighted) ℓ_1 -penalized logistic regression. Let $\check{\gamma}^{(0)}$ and $\check{\beta}^{(0)}$ denote the refitted estimates.
2. Calculate the weights $w(L_i; \hat{\gamma}^{(0)}) = \text{expit}(\hat{\gamma}^{(0)'} L_i) \{1 - \text{expit}(\hat{\gamma}^{(0)'} L_i)\}$, $w(L_i; \hat{\beta}^{(0)}) = \text{expit}(\hat{\beta}^{(0)'} L_i) \{1 - \text{expit}(\hat{\beta}^{(0)'} L_i)\}$, $w(L_i; \check{\gamma}^{(0)})$ and $w(L_i; \check{\beta}^{(0)})$. Calculate the objective function

$$\begin{aligned} \check{\nu}^{(0)} = & \frac{1}{n} \sum_{i=1}^n \log\{1 + \exp(\check{\gamma}^{(0)'} L_i)\} - A_i(\check{\gamma}^{(0)'} L_i) + \log\{1 + \exp(\check{\beta}^{(0)'} L_i)\} \\ & - Y_i(\check{\beta}^{(0)'} L_i) \end{aligned}$$

3. Set $j = 0$ and carry out the following recursive algorithm:

- (a) Set $j = j + 1$.
- (b) Using the initial estimates, re-estimate γ and β as the solutions $\hat{\gamma}^{(j)}$ and $\hat{\beta}^{(j)}$ to

$$\begin{aligned} 0 &= \sum_{i=1}^n w(L_i; \hat{\beta}^{(j-1)'}) \{A_i - \text{expit}(\gamma^T L_i)\} L_i + \lambda_\gamma \delta |\gamma|^{\delta-1} \circ \text{sign}(\gamma) \\ 0 &= \sum_{i=1}^n w(L_i; \hat{\gamma}^{(j-1)'}) \{Y_i - \text{expit}(\beta^T L_i)\} L_i + \lambda_\beta \delta |\beta|^{\delta-1} \circ \text{sign}(\beta) \end{aligned}$$

Similarly, using $w(L_i; \check{\gamma}^{(j-1)'})$ and $w(L_i; \check{\beta}^{(j-1)'})$, obtain the refitted $\check{\gamma}^{(j)}$ and $\check{\beta}^{(j)}$.

- (c) Re-evaluate the objective function as:

$$\begin{aligned} \check{\nu}^{(j)} = & \frac{1}{n} \sum_{i=1}^n \left[\log\{1 + \exp(\check{\gamma}^{(j)'} L_i)\} - A_i(\check{\gamma}^{(j)'} L_i) \right] w(L_i; \check{\beta}^{(j-1)'}) \\ & + \left[\log\{1 + \exp(\check{\beta}^{(j)'} L_i)\} - Y_i(\check{\beta}^{(j)'} L_i) \right] w(L_i; \check{\gamma}^{(j-1)'}) \end{aligned}$$

- (d) If $|\check{\nu}^{(j)} - \check{\nu}^{(j-1)}| < 0.0001$, stop the algorithm, and set $\check{\gamma}_{BR} = \check{\gamma}^{(j)}$ and $\check{\beta}_{BR} = \check{\beta}^{(j)}$.
-

Y_i was normally distributed with mean $1 + \beta^T (|L_{\cdot, [1:3]}|; L_{\cdot, [4:p]})$ and standard deviation

$$\sqrt{\left(\frac{\{1 + \beta^T (|L_{\cdot, [1:3]}|; L_{\cdot, [4:p]})\}^2}{E_n[\{1 + \beta^T (|L_{\cdot, [1:3]}|; L_{\cdot, [4:p]})\}^2]} \right)}.$$

some brief discussion of the results would be helpful.

Table 1: Type I errors based on 1,000 replications in settings with a denser propensity score: $\Sigma = \mathbf{I}_{p \times p}$.

<i>Correct models</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.520	0.813	0.289	0.602
Standard naïve (not forced)	0.272	0.522	0.171	0.336
PDS (pre-specified)	0.526	0.779	0.523	0.517
PO (pre-specified)	0.517	0.758	0.521	0.502
PDS (CV)	0.074	0.070	0.066	0.079
PMLE-DR	0.043	0.064	0.039	0.065
BR-DR	0.054	0.061	0.064	0.054
<i>Incorrect outcome model</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.363	0.576	0.205	0.456
Standard naïve (not forced)	0.169	0.304	0.114	0.196
PDS (pre-specified)	0.344	0.611	0.342	0.310
PO (pre-specified)	0.338	0.596	0.325	0.309
PDS (CV)	0.064	0.068	0.068	0.073
PMLE-DR	0.041	0.047	0.039	0.044
BR-DR	0.055	0.052	0.069	0.041

Table 2: Type I errors based on 1,000 replications in settings with a denser propensity score: $\Sigma = [\sigma_{i,j}]_{1 \leq i,j \leq p}$.

<i>Correct models</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.431	0.368	0.197	0.505
Standard naïve (not forced)	0.173	0.174	0.088	0.203
PDS (pre-specified)	0.132	0.082	0.096	0.143
PO (pre-specified)	0.102	0.071	0.078	0.094
PDS (CV)	0.056	0.071	0.064	0.068
PMLE-DR	0.015	0.006	0.015	0.016
BR-DR	0.047	0.041	0.043	0.041
<i>Incorrect outcome model</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.315	0.265	0.156	0.353
Standard naïve (not forced)	0.126	0.116	0.090	0.137
PDS (pre-specified)	0.097	0.066	0.070	0.091
PO (pre-specified)	0.073	0.066	0.056	0.074
PDS (CV)	0.063	0.060	0.060	0.068
PMLE-DR	0.016	0.022	0.021	0.019
BR-DR	0.038	0.043	0.048	0.037

any thoughts why these rates are too low?

References

- Avagyan, V. and Vansteelandt, S. (2017). Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. *arXiv:1708.03787 [stat]*. arXiv: 1708.03787.
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6):2369–2429.

Table 3: Type I errors based on 1,000 replications in settings with heteroscedastic errors: $\Sigma = \mathbf{I}_{p \times p}$.

<i>Correct models</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.532	0.815	0.279	0.613
Standard naïve (not forced)	0.260	0.550	0.171	0.324
PDS (pre-specified)	0.543	0.779	0.549	0.531
PO (pre-specified)	0.502	0.738	0.517	0.497
PDS (CV)	0.073	0.060	0.062	0.079
PMLE-DR	0.054	0.076	0.039	0.069
BR-DR	0.060	0.062	0.069	0.054

<i>Incorrect outcome model</i>				
	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.375	0.590	0.203	0.436
Standard naïve (not forced)	0.152	0.309	0.105	0.199
PDS (pre-specified)	0.348	0.627	0.369	0.345
PO (pre-specified)	0.310	0.594	0.348	0.322
PDS (CV)	0.066	0.064	0.071	0.068
PMLE-DR	0.040	0.051	0.036	0.056
BR-DR	0.042	0.069	0.072	0.049

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-Selection Inference for Generalized Linear Models With Many Controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg.
- De la Peña, V., Lai, T. L., and Shao, Q.-M. (2009). *Self-normalized processes: limit theory and statistical applications*. Probability and its applications. Springer, Berlin. OCLC: ocn244765605.
- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.
- Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.
- van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.
- von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the r th Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.