Supplementary Appendix to "Testing Mediation Effects Using Logic of Boolean Matrices"

Chengchun Shi and Lexin Li

London School of Economics and Political Science
and University of California at Berkeley

We first outline a multi-split version of our test. We then introduce a DAG learning procedure under a weaker constant variance condition. We next present some additional regularity conditions, followed by two supporting lemmas, then the proofs of the main theoretical results in the paper. Finally, we present some additional numerical results.

We employ the following notation. For any sequence $\{a_n : n \geq 1\}$, $a_n = O(1)$ means $|a_n| \leq C$ for some constant $C > 0$, and $a_n = o(1)$ means $\lim_n a_n = o(1)$. For a sequence of random variables $\{Z_n : n \geq 1\}$, $Z_n = O_p(1)$ means, for any sufficiently small $\varepsilon > 0$, there exists some constant $M > 0$ such that $\Pr(|Z_n| \leq M) \geq 1 - \varepsilon$, and $Z_n = o_p(1)$ means $\{Z_n : n \geq 1\}$ converges in probability to zero. Without loss of generality, we assume $\boldsymbol{\mu}_0 = 0$. To simplify the presentation, we only consider the case where $\widehat{\boldsymbol{\mu}} = 0$ and hence $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$ for $i = 1, \ldots, n$. In the case where $\widehat{\boldsymbol{\mu}} \neq 0$, the theories can be similarly proved.

S1 A multi-split version of the test

We first develop a version of our individual mediator test based on multiple binary splits. This helps improve the power when the sample size is limited, and also helps mitigate the randomization arising from a single binary split. The main idea is to apply the single-split method in Algorithm 1 multiple times, then combine the p-values from all splits. Specifically, we carry out the binary split S times. For the sth binary split, we divide $\{1, \ldots, n\}$ into two disjoint subsets $\mathcal{I}_{s,1} \cup \mathcal{I}_{s,2}$ of equal sizes. We then apply Algorithm 1 to compute the p-values for $H_0(0, q)$ and $H_0(q, d+1)$ for each half of the data. Denote the obtained p-values by $\widehat{p}^{(s,1)}(0, q)$, $\widehat{p}^{(s,1)}(q, d+1)$, $\widehat{p}^{(s,2)}(0, q)$ and $\widehat{p}^{(s,2)}(q, d+1)$, respectively. We next combine these p-values following the idea of Meinshausen et al. (2009), by defining

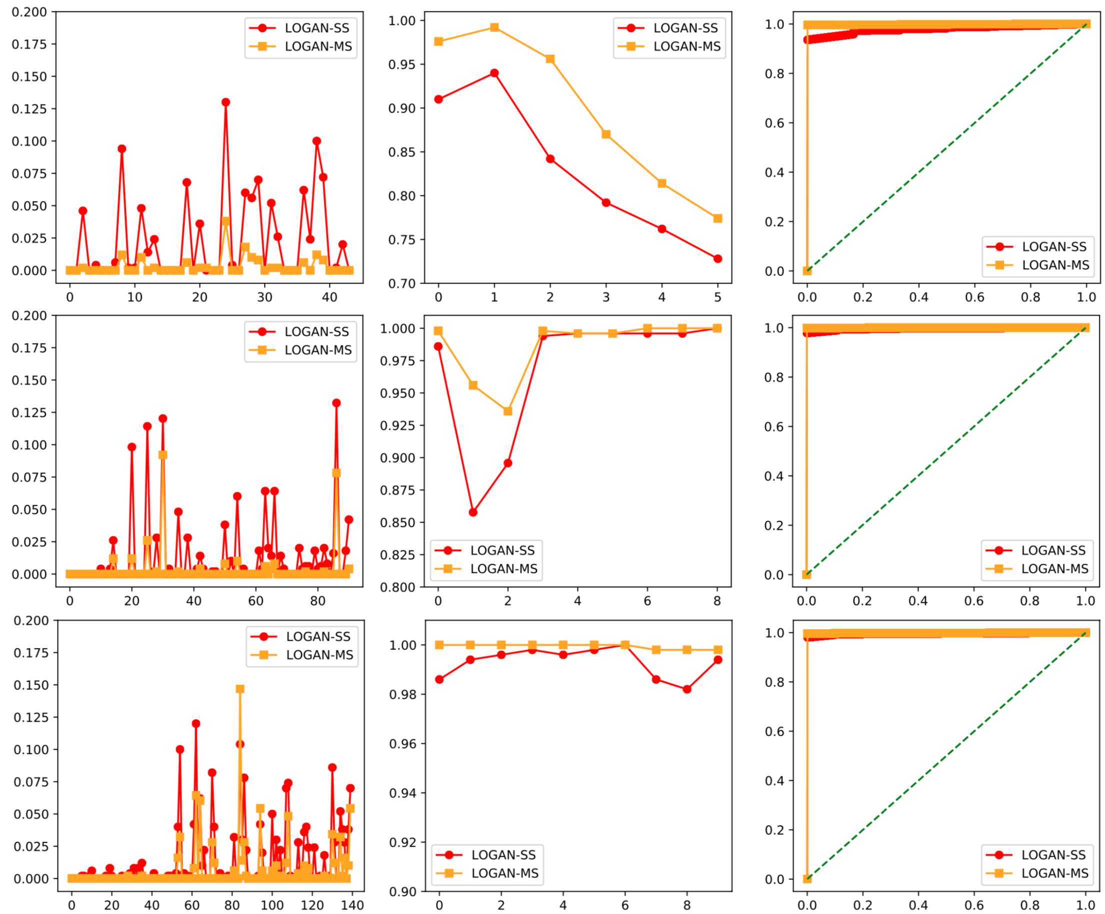

Figure S1: Empirical rejection rate of the single-split (LOGAN-SS) and the multi-split (LOGAN-MS) method. The upper panels: $(d, n) = (50, 100)$, the middle panels: $(d, n) = (100, 250)$, and the bottom panels: $(d, n) = (150, 250)$. The left panels: under H_0, the middle panels: under H_1, where the horizontal axis is the mediator index, and the right panels: the average ROC curve.

$$\widehat{p}(0, q) = \min\left(1, q_\gamma \left[\{\gamma^{-1}\widehat{p}^{(s,\ell)}(0, q), s = 1, \ldots, S, \ell = 1, 2\}\right]\right),$$
$$\widehat{p}(q, d+1) = \min\left(1, q_\gamma \left[\{\gamma^{-1}\widehat{p}^{(s,\ell)}(q, d+1), s = 1, \ldots, S, \ell = 1, 2\}\right]\right),$$

where γ is some constant between 0 and 1, and q_γ is the empirical γ-quantile. In our simulations, we have experimented with a range of values of γ between 0.1 and 0.2, and the results are similar, and thus we set $\gamma = 0.15$ in our implementation. The corresponding p-value for $H_0(q)$ is given by $\max\{\widehat{p}(0, q), \widehat{p}(q, d+1)\}$, following the union-intersection principle.

We apply this multi-split method to the simulation examples in Section 6, and compare with the single-split method. Figure S1 shows the empirical rejection rates of the two methods. It is seen that the multi-split method in general improves over the single-split method, by achieving smaller type-I errors and larger powers.

S2 Learning DAG under a weaker constant variance condition

We note that the constant variance condition in (A3) can be relaxed as follows.

(A3*) The errors ε_i, $i = 0, 1, \ldots, d+1$, are jointly normally distributed and independent. In addition, the error variances $\sigma_i^2 = \mathrm{Var}(\varepsilon_i)$, $i = 1, \ldots, d$, are constant; i.e., $\sigma_1^2 = \cdots = \sigma_d^2 = \sigma_*^2$ for some constant $\sigma_* > 0$.

In other words, the constant variance requirement does not have to be imposed on the exposure and outcome variables. Under this weaker requirement, we first need to ensure $\boldsymbol{W}_0$ remains identifiable, since Peters and Bühlmann (2014, Theorem 1) is no longer directly applicable. In addition, we need to modify the initial estimator of $\boldsymbol{W}_0$ accordingly.

The next lemma shows that $\boldsymbol{W}_0$ remains identifiable under (A3*).

Lemma 3. *Suppose (A1), (A2) and (A3*) hold. Then $\boldsymbol{W}_0$ is identifiable from the joint distribution function of $\boldsymbol{X}$.*

Proof: By (A2), we decompose $\boldsymbol{W}_0$ as,

$$\boldsymbol{W}_0 = \begin{pmatrix} 0 & \mathbf{0}_d^\top & 0 \\ \boldsymbol{W}_{0,1} & \boldsymbol{W}_{1,1} & \mathbf{0}_d \\ W_{0,2} & \boldsymbol{W}_{1,2}^\top & 0 \end{pmatrix}, \tag{S1}$$

where $W_{0,2} \in \mathbb{R}$, $\boldsymbol{W}_{0,1}, \boldsymbol{W}_{1,2} \in \mathbb{R}^d$, $\boldsymbol{W}_{1,1} \in \mathbb{R}^{d\times d}$, and the matrix $\boldsymbol{W}_{1,1}$ is acyclic under (A1).

We first note that $W_{0,2}$ and $\boldsymbol{W}_{1,2}$ correspond to the regression coefficients of $(E, \boldsymbol{M})^\top$ on Y. Under the given model, the covariance matrix of $(E, \boldsymbol{M}^\top)$ is non-degenerated. As such, $W_{0,2}$ and $\boldsymbol{W}_{1,2}$ are uniquely determined by the distribution function of $\boldsymbol{X}$.

We next show that $\boldsymbol{W}_{1,1}$ is also uniquely determined by the distribution function of $\boldsymbol{X}$. For each $j = 1, \ldots, d$, let $\widetilde{M}_j$ denote the population residual adjusted by the exposure, i.e., $\widetilde{M}_j = (M_j - \mu_{0,j}) - \mathrm{corr}(M_j, E)(E - \mu_{0,0})$. It follows that the set of residuals $\widetilde{\boldsymbol{M}}$ satisfy that $\widetilde{\boldsymbol{M}} = \boldsymbol{W}_{1,1}\widetilde{\boldsymbol{M}} + (\varepsilon_1, \ldots, \varepsilon_d)^\top$. By (A1), $\boldsymbol{W}_{1,1}$ is acyclic. As such, $\widetilde{\boldsymbol{M}}$ forms a structural linear equation with the coefficient matrix $\boldsymbol{W}_{1,1}$. Under (A3*), all the residuals have the constant variance. It then follows from Theorem 1 of Peters and Bühlmann (2014) that $\boldsymbol{W}_{1,1}$ is identifiable.

Finally, we note that $\boldsymbol{W}_{0,1}$ satisfies $\boldsymbol{W}_{0,1} = \mathrm{cov}(E, \boldsymbol{M} - \boldsymbol{W}_{1,1}\boldsymbol{M})/\mathrm{Var}(E)$. It follows from the identifiability of $\boldsymbol{W}_{1,1}$ that $\boldsymbol{W}_{0,1}$ is identifiable as well.

This completes the proof. □

We next outline a modified initial DAG estimation procedure under the new constant variance condition (A3*). Similar to Proposition 1, we can show that this new estimator satisfies the oracle inequality as well.

Following the decomposition of $\boldsymbol{W}_0$ in (S1), we first estimate $\boldsymbol{W}_{0,2}$ and $\boldsymbol{W}_{1,2}$ using penalized regressions of $(E, \boldsymbol{M})^\top$ on Y such as MCP, LASSO, SCAD, and Dantzig selector. Denote the corresponding estimators as $\widetilde{W}_{0,2}^{(\ell)}$ and $\widetilde{\boldsymbol{W}}_{1,2}^{(\ell)}$.

We next estimate $\boldsymbol{W}_{1,1}$, by first regressing $\boldsymbol{M}_i$ on E_i, $i = 1, \ldots, n$, to obtain the estimated residual $\widehat{\widetilde{\boldsymbol{M}}}_i$, then employing the method of Zheng et al. (2018) to solve

$$\widetilde{\boldsymbol{W}}_{1,1}^{(\ell)} = \operatorname*{arg\,min}_{\boldsymbol{W} \in \mathbb{R}^{d\times d}} \sum_{i \in \mathcal{I}_\ell} \|\widehat{\widetilde{\boldsymbol{M}}}_i - \boldsymbol{W} \widehat{\widetilde{\boldsymbol{M}}}_i\|_2^2 + \lambda |\mathcal{I}_\ell| \sum_{i,j} |W_{i,j}| \text{ subject to } \operatorname{trace}\{\exp(\boldsymbol{W} \circ \boldsymbol{W})\} = d.$$

After obtaining $\widetilde{\boldsymbol{W}}_{1,1}^{(\ell)}$, we set $\widetilde{\boldsymbol{W}}_{0,1}^{(\ell)} = \widetilde{\boldsymbol{W}}_{1,1}^{(\ell)} \widehat{\operatorname{cov}}(\boldsymbol{M}, E) \widehat{\operatorname{Var}}^{-1}(E)$, where $\widehat{\operatorname{cov}}(\boldsymbol{M}, E)$ is the sampling covariance estimator of $\boldsymbol{M}$, and $\widehat{\operatorname{Var}}(E)$ is the sampling variance estimator of E.

Finally, we put together $\widetilde{\boldsymbol{W}}_{0,1}^{(\ell)}$, $\widetilde{\boldsymbol{W}}_{1,1}^{(\ell)}$, $\widetilde{W}_{0,2}^{(\ell)}$ and $\widetilde{\boldsymbol{W}}_{1,2}^{(\ell)}$ according to (S1) to form the modified initial estimator $\widetilde{\boldsymbol{W}}^{(\ell)}$ for $\boldsymbol{W}_0$.

S3 Additional regularity conditions

We introduce two additional regularity conditions for the theoretical guarantees of the proposed test. We begin with some notation. Define the limit of the estimator $\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2)$

$$\boldsymbol{\beta}_0^{(\ell)}(j_1, j_2) = \operatorname*{arg\,min}_{\boldsymbol{\beta}: \beta_{j_2}=0, \operatorname{supp}(\boldsymbol{\beta}) \in \operatorname{ACT}(j_1, \widetilde{\boldsymbol{W}}^{(\ell)})} \operatorname{E}\left(X_{j_2} - \boldsymbol{\beta}^\top \boldsymbol{X}_i\right)^2.$$

Any permutation $\pi = (\pi_0, \pi_1, \ldots, \pi_d, \pi_{d+1})^\top$ of $\{0, 1, \ldots, d, d+1\}$ determines an order of the mediators $\{M_j\}_{1 \le j \le d}$. Define

$$\boldsymbol{W}_{\pi_j}(\pi) = \operatorname*{arg\,min}_{\boldsymbol{\beta}: \operatorname{supp}(\boldsymbol{\beta}) \in \{\pi_0, \pi_1, \ldots, \pi_{j-1}, \pi_j\}} \operatorname{E}\left(M_{\pi_j} - \boldsymbol{\beta}^\top \boldsymbol{M}\right)^2, \quad \text{for } j = 1, \ldots, d.$$

Let $\boldsymbol{W}(\pi) = \{\boldsymbol{W}_0(\pi), \boldsymbol{W}_1(\pi), \ldots, \boldsymbol{W}_{d+1}(\pi)\}^\top$. It corresponds to the coefficient matrix obtained by doing a Gram-Schmidt orthogonalization, starting with X_{π_0}, and finishing by projecting $X_{\pi_{d+1}}$ on $X_{\pi_0}, X_{\pi_1}, \ldots, X_{\pi_d}$. Let $\boldsymbol{\Omega}(\pi)$ be a diagonal matrix where the diagonal elements $\omega_0^2(\pi), \omega_1^2(\pi), \ldots, \omega_{d+1}^2(\pi)$ correspond to the error variances, $\operatorname{Var}\{X_0 - \boldsymbol{X}^\top \boldsymbol{W}_0(\pi)\}, \operatorname{Var}\{X_1 - \boldsymbol{X}^\top \boldsymbol{W}_1(\pi)\}, \ldots, \operatorname{Var}\{X_{d+1} - \boldsymbol{X}^\top \boldsymbol{W}_{d+1}(\pi)\}$, respectively. Let Π^* denote the set consisting of all true orderings π^* such that $\boldsymbol{W}(\pi^*) = \boldsymbol{W}_0$. Note that π^* may not be unique. As an illustration, consider the DAG in Figure S2, where both $(0, 1, 2, 3)$

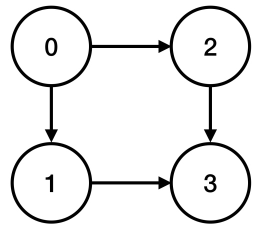

Figure S2: An illustrative DAG with four nodes, where both $(0, 1, 2, 3)$ and $(0, 2, 1, 3)$ correspond to the true orderings.

and $(0, 2, 1, 3)$ correspond to the true orderings of the four nodes.

We impose the following additional regularity conditions. In particular, (A5) is required to establish the consistency and FDR control of the proposed test, while (A6) is to establish the oracle inequality for the initial estimator.

(A5) There exist some constants $\kappa_3, \kappa_4, \kappa_5, \kappa_6 > 0$, with $\kappa_4 + \kappa_5 > 1/2$, such that, $\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\|_2 \leq \kappa_3 n^{-\kappa_4}$, $\|\overline{\boldsymbol{W}}_{j_1}^{(\ell)} - \boldsymbol{W}_{0,j_1}\|_2 \leq \kappa_3 n^{-\kappa_5}$ and $\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\|_1 \leq \kappa_3 n^{-\kappa_6}$, $\|\overline{\boldsymbol{W}}_{j_1}^{(\ell)} - \boldsymbol{W}_{0,j_1}\|_1 \leq \kappa_3 n^{-\kappa_6}$, for any $0 \leq j_1, j_2 \leq d+1, \ell = 1, 2$.

(A6) There exists a constant $\omega > 0$ such that for all $\pi \notin \Pi^*$, $\frac{1}{d}\sum_{j=0}^{d+1}\left\{\omega_j^2(\pi) - \sigma_*^2\right\}^2 > \omega$.

Condition (A5) is mild. This is because, when (A4) holds and $\boldsymbol{\beta}_0^{(\ell)}$ is estimated via the MCP, LASSO, SCAD, or Dantzig selector, we have $\max_{j_1,j_2,\ell}\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\|_2 \leq O(1)n^{-1/2}\sqrt{s^* \log n}$ and $\max_{j_1,j_2,\ell}\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\|_1 \leq O(1)n^{-1/2}s^*\sqrt{\log n}$, with probability approaching one, where $s^* = \max_{j_1,j_2,\ell}|\mathcal{M}^{(\ell)}(j_1, j_2)|$ denotes the maximum sparsity size, $\mathcal{M}^{(\ell)}(j_1, j_2) = \operatorname{supp}\left\{\boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\right\}$, and $O(1)$ denotes some positive constant. Similarly, we have $\max_{j,\ell}\|\overline{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_2 \leq O(1)n^{-1/2}\sqrt{s_0 \log n}$ and $\max_{j,\ell}\|\overline{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_1 \leq O(1)n^{-1/2}s_0\sqrt{\log n}$, with probability approaching one. Therefore, Condition (A5) holds as long as $s_0, s^* = O(n^{\kappa_7})$ for some $\kappa_7 < 1/2$.

Condition (A6) is referred to as the "omega-min" condition in van de Geer and Bühlmann (2013). It essentially guarantees that the true ordering of the mediators can be consistently estimated, which is needed to establish the oracle inequality for the estimator from (11) of Zheng et al. (2018). When the number of mediators d is fixed and the error variances are equal as in (A3), this condition automatically holds.

Finally, we make some remark on why the usual debiasing strategy may not be directly applicable to relax the regularity condition (A4) in Section 5.1 of the paper. Specifically, if (A4) does not hold and the true ordering cannot not be recovered, then no matter whether we debias the estimated coefficient matrix or not, the resulting estimator for $\boldsymbol{W}_0$ may not be consistent. We illustrate with a simple example. Consider a DAG with two variables, where $X_1 = \varepsilon_1$, $X_2 = aX_1 + \varepsilon_2$ for some $a \neq 0$, and ε_1 and ε_2 are independent mean-zero random errors. Then the corresponding coefficient matrix is

$$\boldsymbol{W}_0 = \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix}.$$

Meanwhile, we note that the linear structure equation can be rewritten as $X_2 = \varepsilon_2^*$ and $X_1 = a(a^2+1)^{-1}X_2 + \varepsilon_1^*$, for some mean-zero random errors ε_1^* and ε_2^*. In addition, with some calculation, we have that

$$\begin{aligned} \varepsilon_2^* &= a\varepsilon_1 + \varepsilon_2, \\ \varepsilon_1^* &= \frac{1}{a^2+1}\varepsilon_1 - \frac{a}{a^2+1}\varepsilon_2, \end{aligned}$$

and thus ε_1^* and ε_2^* are independent. Then the corresponding coefficient matrix $\boldsymbol{W}_0^*$ becomes

$$\begin{pmatrix} 0 & a(a^2+1)^{-1} \\ 0 & 0 \end{pmatrix}.$$

If the ordering is not correctly specified and (A4) does not hold, then debiasing or not, the second element on the first row will be close to $a(a^2+1)^{-1}$, rather than the true value 0. As such, the debiased estimator is not consistent when (A4) is violated. This simple example reflects the challenge of post-selection inference in our setting.

S4 Supporting lemmas

Next, we present two supporting lemmas. Lemma 4 establishes the convergence rate of the variance estimator $\widehat{\sigma}_*^2$, whereas Lemma 5 is needed for the limiting distribution.

Lemma 4. *Suppose (A5) holds, and $\|\boldsymbol{W}_0\|_2$ is bounded. Then $|\widehat{\sigma}_*^2 - \sigma_*^2| = O(n^{-\kappa_7})$ for some constant κ_7 satisfying $0 < \kappa_7 \leq \min(2\kappa_5, 1/2)$, with probability tending to 1.*

Proof: It suffices to show that, for $\ell = 1, 2$ and any κ_7 satisfying that $\kappa_7 \leq 2\kappa_5$, $\kappa_7 < 1/2$,

$$\frac{1}{(d+2)|\mathcal{I}_\ell^c|}\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell^c}|X_{i,j} - \overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}_i|_2^2 - \sigma_*^2 = O(n^{-\kappa_7}), \tag{S2}$$

with probability tending to 1. In turn, it suffices to show that,

$$\Pr\left(\left|\frac{1}{(d+2)|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\sum_{j=0}^{d+1}|X_{i,j} - \overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}_i|_2^2 - \sigma_*^2\right| > 2\kappa_3^2 n^{-\kappa_7} \;\middle|\; \overline{\boldsymbol{W}}^{(\ell)}\right) = o_p(1). \tag{S3}$$

This is because, if (S3) holds, by bounded convergence theorem, we have

$$\Pr\left(\left|\frac{1}{(d+2)|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\sum_{j=0}^{d+1}|X_{i,j}-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}_i|_2^2-\sigma_*^2\right|>2\kappa_3^2 n^{-\kappa_7}\right)=o(1),$$

which in turn yields (S2).

For the conditional mean of the left-hand-side of (S2) given $\{\boldsymbol{X}_i : i \in \mathcal{I}_\ell\}$,

$$\begin{aligned}
&\frac{1}{(d+2)}\mathrm{E}\left(\sum_{j=0}^{d+1}|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|_2^2 \mid \overline{\boldsymbol{W}}^{(\ell)}\right)-\sigma_*^2\\
=\;&\frac{1}{(d+2)}\sum_{j=0}^{d+1}\mathrm{E}\|X_j-\boldsymbol{W}_{0,j}^\top\boldsymbol{X}\|_2^2-\sigma_*^2+\frac{1}{(d+2)}\|\boldsymbol{W}_0-\overline{\boldsymbol{W}}^{(\ell)}\|_2^2\\
=\;&\frac{1}{(d+2)}\|\boldsymbol{W}_0-\overline{\boldsymbol{W}}^{(\ell)}\|_2^2=\frac{1}{(d+2)}\sum_{j=0}^{d+1}\|\boldsymbol{W}_{0,j}-\overline{\boldsymbol{W}}_j^{(\ell)}\|_2^2\;\le\;\kappa_3^2 n^{-2\kappa_5}\;\le\;\kappa_3^2 n^{-\kappa_7},
\end{aligned}$$

where the first equality is due to the fact that $\mathrm{E}\boldsymbol{X}=0$ and the second-to-last inequality is due to Condition (A5). The event defined in (S3) occurs only when

$$\left|\frac{1}{(d+2)|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\sum_{j=0}^{d+1}|X_{i,j}-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}_i|_2^2-\frac{1}{(d+2)}\mathrm{E}\left(\sum_{j=0}^{d+1}|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|_2^2 \mid \overline{\boldsymbol{W}}^{(\ell)}\right)\right|>\kappa_3^2 n^{-\kappa_7}.$$

Thus, to prove (S3), it suffices to show that

$$\begin{aligned}
\Pr\Bigg(\Bigg|\frac{1}{(d+2)|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\sum_{j=0}^{d+1}|X_{i,j}-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}_i|_2^2-\frac{1}{(d+2)}\mathrm{E}\left(\sum_{j=0}^{d+1}|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|_2^2 \mid \overline{\boldsymbol{W}}^{(\ell)}\right)\Bigg|&\\
>\kappa_3^2 n^{-\kappa_7}|\overline{\boldsymbol{W}}^{(\ell)}\Bigg)&=o_p(1).
\end{aligned}$$

By Chebyshev's inequality, this probability is bounded from above by

$$\frac{n^{2\kappa_7-1}}{2\kappa_3^4(d+2)^2}\mathrm{Var}\left(\sum_{j=0}^{d+1}|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|^2 \mid \overline{\boldsymbol{W}}^{(\ell)}\right)\le\frac{n^{2\kappa_7-1}}{\kappa_3^4(d+2)^2}\mathrm{E}\left(\left\{\sum_{j=0}^{d+1}\|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|_2^2\right\}^2 \,\middle|\, \overline{\boldsymbol{W}}^{(\ell)}\right).$$

Since $\kappa_7 < 1/2$, it suffices to show that, with probability approaching one,

$$\frac{1}{(d+2)^2}\mathrm{E}\left(\left\{\sum_{j=0}^{d+1}|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|_2^2\right\}^2 \,\middle|\, \overline{\boldsymbol{W}}^{(\ell)}\right)=O(1), \tag{S4}$$

By Cauchy-Schwarz inequality, the left-hand-side of (S4) is bounded from above by

$$\frac{1}{d+2}\sum_{j=0}^{d+1}\mathrm{E}\left\{|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|^4\mid\overline{\boldsymbol{W}}^{(\ell)}\right\}\leq$$
$$\frac{16}{d+2}\sum_{j=0}^{d+1}\left(\mathrm{E}\left\{|X_j|^4\mid\overline{\boldsymbol{W}}^{(\ell)}\right\}+\mathrm{E}\left\{|\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|^4\mid\overline{\boldsymbol{W}}^{(\ell)}\right\}\right). \tag{S5}$$

For any random variable Z and any constant $\kappa>0$, by Taylor's theorem, we have

$$\mathrm{E}\left\{\exp\left(\frac{Z^2}{\kappa^2}\right)\right\}=\sum_{k=0}^{+\infty}\frac{\mathrm{E}\left(Z^{2k}\right)}{\kappa^{2k}k!}.$$

It follows that $\mathrm{E}\{Z^4/(2\kappa^4)\}\leq\mathrm{E}\{\exp(Z^2/\kappa^2)\}-1$. Let $\|Z\|_{\psi_p}$ denote its Orlicz norm, i.e.,

$$\|Z\|_{\psi_p}=\inf\left[C>0:\mathrm{E}\left\{\exp\left(\frac{|Z|^p}{|C|^p}\leq 2\right)\right\}\right].$$

By definition, we have $\mathrm{E}\{Z^4/(2\|Z\|_{\psi_2}^4)\}\leq 1$, and henceforth, $\mathrm{E}(Z^4)\leq 2\|Z\|_{\psi_2}^4$.

Under our model assumptions, the covariance matrix $\boldsymbol{\Sigma}_0=\mathrm{cov}(\boldsymbol{X})$ is given by $\sigma_*^2(\boldsymbol{I}_{d+2}-\boldsymbol{W}_0)^{-1}\{(\boldsymbol{I}_{d+2}-\boldsymbol{W}_0)^{-1}\}^\top$, where $\boldsymbol{I}_{d+2}$ is a $(d+2)\times(d+2)$ identity matrix. For any $\boldsymbol{a}\in\mathbb{R}^{d+2}$, by Cauchy-Schwarz inequality,

$$\boldsymbol{a}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{a}\leq\kappa_*\|\boldsymbol{a}^\top(\boldsymbol{I}_{d+2}-\boldsymbol{W}_0)\|_2^2\leq\kappa_*\|\boldsymbol{a}\|_2^2\|\boldsymbol{I}_{d+2}-\boldsymbol{W}_0\|_2^2\leq 2\kappa_*\|\boldsymbol{a}\|_2^2\left(\|\boldsymbol{I}_{d+2}\|_2^2+\|\boldsymbol{W}_0\|_2^2\right),$$

for some constant $\kappa_*>0$. Since $\|\boldsymbol{W}_0\|_2$ is bounded, it implies that the maximum eigenvalues of $\boldsymbol{\Sigma}_0^{-1}$ is bounded. Thus, the minimum eigenvalue of $\boldsymbol{\Sigma}_0$ is bounded away from zero. Also, note that $(\boldsymbol{I}_{d+2}-\boldsymbol{W}_0)^{-1}=\boldsymbol{I}_{d+2}+\boldsymbol{W}_0$. Following similar arguments, the maximum eigenvalue of $\boldsymbol{\Sigma}_0$ is bounded as well. Since $\boldsymbol{X}$ is jointly normal with bounded $\lambda_{\max}(\Sigma_0)$, we have, for some constant $\kappa>0$ and any vector $\boldsymbol{a}\in\mathbb{R}^{d+2}$,

$$\|\boldsymbol{a}^\top\boldsymbol{X}\|_{\psi_2}\leq\kappa\|\boldsymbol{a}\|_2. \tag{S6}$$

It then follows from (S6) that conditional on $\overline{\boldsymbol{W}}^{(\ell)}$,

$$\mathrm{E}\left(|X_j|^4\right)\leq 2\|X_j\|_{\psi_2}^4\leq 2\kappa^4,\quad\text{and}\quad\mathrm{E}\left(|\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|^4\right)\leq 2\|\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}\|_{\psi_2}^4\leq 2\kappa^4\|\overline{\boldsymbol{W}}_j^{(\ell)}\|_2^4,$$

for any $j=1,\dots,d$. It follows from (S5) that conditional on $\overline{\boldsymbol{W}}^{(\ell)}$,

$$\frac{1}{(d+2)}\sum_{j=1}^{d}\mathrm{E}\left(|X_j-\overline{\boldsymbol{W}}_j^{(\ell)\top}\boldsymbol{X}|^4\right)\leq\frac{32\kappa^4}{d+2}\sum_{j=0}^{d+1}\left(1+\|\overline{\boldsymbol{W}}_j^{(\ell)}\|_2^4\right)\leq 32\kappa^4\left(1+\max_j\|\overline{\boldsymbol{W}}_j^{(\ell)}\|_2^4\right).$$

By (A5), we have $\max_j \|\overline{\boldsymbol{W}}_j^{(\ell)}\|_2 \leq \max_j \|\boldsymbol{W}_{0,j}\|_2 + \max_j \|\boldsymbol{W}_{0,j} - \overline{\boldsymbol{W}}_j^{(\ell)}\|_2 \leq \max_j \|\boldsymbol{W}_{0,j}\|_2 + \kappa_3 n^{-\kappa_5}$. Since $\|\boldsymbol{W}_0\|_2$ is bounded, we have $\|\boldsymbol{e}_j^\top \boldsymbol{W}_0\|_2 = O(1)$, and hence $\|\boldsymbol{W}_{0,j}\|_2 = O(1)$, where $\boldsymbol{e}_j$ denotes a $(d+2)$-dimensional vector with the jth element equal to one and the rest equal to zero. It follows that $\sum_{j=0}^{d+1} \mathrm{E}\left\{ |X_j - \overline{\boldsymbol{W}}_j^{(\ell)\top} \boldsymbol{X}|^4/(d+2) \middle| \overline{\boldsymbol{W}}^{(\ell)} \right\} = O(1)$, with probability approaching one. Then (S4) is proven. This completes the proof. $\square$

Lemma 5. *Suppose (A1), (A4) hold. Then $ACT(j, \widetilde{\boldsymbol{W}}^{(\ell)})$ contains no descendants of j.*

Proof: Suppose there exists some $j' \in \mathrm{ACT}(j, \widetilde{\boldsymbol{W}}^{(\ell)})$, such that j' is a descendant of j. By definition, there exists a directed path from X_j to $X_{j'}$: $X_j \to X_{i_1} \to \dots X_{i_K} \to X_{j'}$. By Condition (A4), we have $j \in \mathrm{ACT}(i_1, \widetilde{\boldsymbol{W}}^{(\ell)})$, $i_k \in \mathrm{ACT}(i_{k+1}, \widetilde{\boldsymbol{W}}^{(\ell)})$, for $k = 1, \dots, K-1$ and $i_K \in \mathrm{ACT}(j', \widetilde{\boldsymbol{W}}^{(\ell)})$. This, together with $j' \in \mathrm{ACT}(j, \widetilde{\boldsymbol{W}}^{(\ell)})$, implies that there exists a directed path from X_j to X_j on the DAG generated by $\widetilde{\boldsymbol{W}}^{(\ell)}$. Then the acyclic constraint of $\widetilde{\boldsymbol{W}}^{(\ell)}$ is violated. This completes the proof. $\square$

S5 Proof of Lemma 2

To prove Lemma 2, it suffices to show that

$$(|\boldsymbol{W}_0|^{(k)})_{q_2,q_1} = \max_{0\leq j_1,\dots,j_{k-1}\leq d+1} \min\left(|W_{0,j_1,q_1}|, \min_{l\in\{1,\dots,k-2\}} |W_{0,j_{l+1},j_l}|, |W_{0,q_2,j_{k-1}}| \right). \tag{S7}$$

Lemma 2 can then be similarly proven as Lemma 1. We use induction to prove (S7) for any $q_1 = 0, \dots, d$, and $q_2 = 1, \dots, d+1$. When $k = 2$, by the definition of $\otimes$, we have $(|\boldsymbol{W}_0|^{(2)})_{q_2,q_1} = \max_{0\leq j\leq d+1} \min(|W_{0,j,q_1}|, |W_{0,q_2,j}|)$. Thus, (S7) holds with $k = 2$.

Suppose (S7) holds with $k = t$ for some $t \geq 2$, i.e,

$$(|\boldsymbol{W}_0|^{(t)})_{q_2,q_1} = \max_{0\leq j_1,\dots,j_{t-1}\leq d+1} \min\left(|W_{0,j_1,q_1}|, \min_{l\in\{1,\dots,t-2\}} |W_{0,j_{l+1},j_l}|, |W_{0,q_2,j_{t-1}}| \right). \tag{S8}$$

Therefore,

$$\begin{aligned}
& (|\boldsymbol{W}_0|^{(t+1)})_{q_2,q_1} = (|\boldsymbol{W}_0|^{(t)} \circ |\boldsymbol{W}_0|)_{q_2,q_1} = \max_{j\in\{0,\dots,d+1\}} \min\{(|\boldsymbol{W}_0|^{(t)})_{j,q_1}, |W_{0,q_2,j}|\} \\
= & \max_{j\in\{0,\dots,d+1\}} \min\left\{ \max_{0\leq j_1,\dots,j_{t-1}\leq d+1} \min\left(|W_{0,j_1,q_1}|, \min_{l\in\{1,\dots,t-2\}} |W_{0,j_{l+1},j_l}|, |W_{0,j,j_{t-1}}| \right), |W_{0,q_2,j}| \right\} \\
= & \max_{j\in\{0,\dots,d+1\}} \max_{0\leq j_1,\dots,j_{t-1}\leq d+1} \min\left(|W_{0,j_1,q_1}|, \min_{l\in\{1,\dots,t-2\}} |W_{0,j_{l+1},j_l}|, |W_{0,j,j_{t-1}}|, |W_{0,q_2,j}| \right) \\
= & \max_{0\leq j_1,\dots,j_{t-1},j_t\leq d+1} \min\left(|W_{0,j_1,q_1}|, \min_{l\in\{1,\dots,t-1\}} |W_{0,j_{l+1},j_l}|, |W_{0,q_2,j_t}| \right).
\end{aligned}$$

Thus, (S8) holds with $k = t+1$. This completes the proof of Lemma 2. $\square$

S6 Proof of Theorem 1

Outline of the proof: Our goal is to prove

$$\Pr\left\{\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}\sqrt{|\mathcal{I}_\ell^c|}|\widehat{W}_{i,j}^{(\ell)}-W_{0,i,j}|>\widehat{c}^{(\ell)}(q_1,q_2)\right\}=\frac{\alpha}{2}+o(1), \tag{S9}$$

for any $q_1=0,\dots,d, q_2=1,\dots,d+1$ and $\ell=1,2$. Then the validity of our test follows by the union-intersection principle. We begin with an outline of our proof, which relies on the high-dimensional central limit theorem that was recently developed by Chernozhukov et al. (2013) and Chernozhukov et al. (2014). We divide the proof into three steps.

In Step 1, we show that,

$$\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}\left|\sqrt{|\mathcal{I}_\ell^c|}\left(\widehat{W}_{j_1,j_2}^{(\ell)}-W_{0,j_1,j_2}\right)-\eta_{0,j_1,j_2}^{(\ell)}\right|=o_p(\log^{-1/2}n), \tag{S10}$$

where

$$\eta_{0,j_1,j_2}^{(\ell)}=\frac{|\mathcal{I}_\ell^c|^{-1/2}\sum_{i\in\mathcal{I}_\ell^c}\left\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\varepsilon_{i,j_1}}{\mathrm{E}\left[X_{j_1}\left\{X_{j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}\right\}\mid\widetilde{\boldsymbol{W}}^{(\ell)}\right]}. \tag{S11}$$

This further implies that

$$\underbrace{\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}\sqrt{|\mathcal{I}_\ell^c|}|\widehat{W}_{j_1,j_2}^{(\ell)}-W_{0,j_1,j_2}|}_{\widehat{S}^{(\ell)}}=\underbrace{\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}|\eta_{0,j_1,j_2}^{(\ell)}|}_{S_0^{(\ell)}}+o_p(\log^{-1/2}n). \tag{S12}$$

Let $\mathcal{M}^{(\ell)}(j_1,j_2)$ denote the support of $\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)$. By definition, for any $j\in\mathcal{M}^{(\ell)}(j_1,j_2)$, we have $j,j_2\in\mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})$. Since $\widetilde{\boldsymbol{W}}^{(1)}$ and $\widetilde{\boldsymbol{W}}^{(2)}$ satisfy the acyclic constraint in (A1), by Lemma 5, neither j nor j_2 is a descendant of j_1. As a result, X_{j_1} is conditionally independent of X_{j_2} and X_j given its parents. As such, the numerator of $\eta_{0,j_1,j_2}^{(\ell)}$ forms a sum of i.i.d. mean zero random variables.

In Step 2, we show that,

$$\sup_{z\in\mathbb{R}}\left|\Pr\left(S_0^{(\ell)}\le z|\widetilde{\boldsymbol{W}}^{(\ell)}\right)-\Pr\left(\|N(0,V_0)\|_\infty\le z|\widetilde{\boldsymbol{W}}^{(\ell)}\right)\right|=o(1), \tag{S13}$$

where V_0 is a matrix involving the covariance of $\eta_{0,j_1,j_2}^{(\ell)}$ in (S11) and is defined in Step 2.

In Step 3, we show that, for some constant $\kappa^*>0$,

$$\|V_0-\widehat{V}\|_{\infty,\infty}=O_p(n^{-\kappa^*}), \tag{S14}$$

where $\|\cdot\|_{\infty,\infty}$ denotes the elementwise max-norm, and $\widehat{V}$ is a matrix involving the covariance of $\eta^{*(\ell)}_{j_1,j_2}$ in (16) of the paper, and is defined later in Step 3.

Following (S13) and bounded convergence theorem, we have that,

$$\sup_{z\in\mathbb{R}}\left|\Pr\left(S_0^{(\ell)}\le z\right)-\Pr\left(\|N(0,V_0)\|_\infty\le z\right)\right|=o(1).$$

This, together with (S12), yields that

$$\begin{aligned}\Pr\left(\widehat{S}^{(\ell)}\le z\right)&\ge\Pr\left(\|N(0,V_0)\|_\infty\le z-\varepsilon\log^{-1/2}n\right)-o(1),\\ \Pr\left(\widehat{S}^{(\ell)}\le z\right)&\le\Pr\left(\|N(0,V_0)\|_\infty\le z+\varepsilon\log^{-1/2}n\right)+o(1),\end{aligned}\tag{S15}$$

for any sufficiently small $\varepsilon>0$, where the little-o term is uniform in z. Using similar arguments for (S15) and also Lemma 3.1 of Chernozhukov et al. (2015), we have by (S14)

$$\begin{aligned}\Pr\left(\widehat{S}^{(\ell)}\le z\right)&\ge\Pr\left(\|N(0,\widehat{V})\|_\infty\le z-2\varepsilon\log^{-1/2}n|\widehat{V}\right)-o(1),\\ \Pr\left(\widehat{S}^{(\ell)}\le z\right)&\le\Pr\left(\|N(0,\widehat{V})\|_\infty\le z+2\varepsilon\log^{-1/2}n|\widehat{V}\right)+o(1),\end{aligned}$$

for any sufficiently small $\varepsilon>0$. Set $z=\widehat{c}^{(\ell)}(q_1,q_2)$. Since the little-o term is uniform in $z\in\mathbb{R}$, we have

$$\begin{aligned}\Pr\left\{\widehat{S}^{(\ell)}\le\widehat{c}^{(\ell)}(q_1,q_2)\right\}&\ge\Pr\left\{\|N(0,\widehat{V})\|_\infty\le\widehat{c}^{(\ell)}(q_1,q_2)-2\varepsilon\log^{-1/2}n|\widehat{V}\right\}-o(1),\\ \Pr\left\{\widehat{S}^{(\ell)}\le\widehat{c}^{(\ell)}(q_1,q_2)\right\}&\le\Pr\left\{\|N(0,\widehat{V})\|_\infty\le\widehat{c}^{(\ell)}(q_1,q_2)+2\varepsilon\log^{-1/2}n|\widehat{V}\right\}+o(1).\end{aligned}\tag{S16}$$

We show in Step 2 that all diagonal elements in V_0 are well bounded away from zero. Henceforth, with probability approaching 1, all diagonal elements in $\widehat{V}$ are well bounded away from zero as well. It follows from Theorem 1 of Chernozhukov et al. (2017) that

$$\begin{aligned}\Pr\left\{\|N(0,\widehat{V})\|_\infty\le\widehat{c}^{(\ell)}(q_1,q_2)+2\varepsilon\log^{-1/2}n|\widehat{V}\right\}-\Pr\left\{\|N(0,\widehat{V})\|_\infty\le\widehat{c}^{(\ell)}(q_1,q_2)\right.\\ \left.-2\varepsilon\log^{-1/2}n|\widehat{V}\right\}\le O(1)\varepsilon\log^{1/2}d\log^{-1/2}n,\end{aligned}$$

where $O(1)$ denotes some positive constant. Under the given conditions, we have $\log d=O(\log n)$. The right-hand-side is bounded by $\kappa\varepsilon$ for some constant $\kappa>0$. This, together with (S16), yields that

$$\left|\Pr\left\{\widehat{S}^{(\ell)}\le\widehat{c}^{(\ell)}(q_1,q_2)\right\}-\Pr\left\{\|N(0,\widehat{V})\|_\infty\le\widehat{c}^{(\ell)}(q_1,q_2)|\widehat{V}\right\}\right|\le\kappa\varepsilon+o(1).$$

Since ε can be made arbitrarily small, (S9) follows, which completes the proof of Theorem 1. Next, we give detailed proofs for each step.

Step 1: The proof of this step relies on the arguments developed to establish the limiting distribution of the debiased LASSO (van de Geer et al., 2014) and the decorrelated score statistic (Ning and Liu, 2017). We aim to establish the upper bound for $\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}$ $|\sqrt{|\mathcal{I}_\ell^c|}(\widehat{W}_{j_1,j_2}-W_{0,j_1,j_2})-\eta^{(\ell)}_{j_1,j_2}|$, and for $\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}|\eta^{(\ell)}_{j_1,j_2}-\eta^{(\ell)}_{0,j_1,j_2}|$. Together, these two upper bounds would lead to (S10).

To obtain the first upper bound, we have that,

$$
\begin{aligned}
&\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}\left|\sqrt{|\mathcal{I}_\ell^c|}\left(\widehat{W}_{j_1,j_2}-W_{0,j_1,j_2}\right)-\eta^{(\ell)}_{j_1,j_2}\right| \\
\leq\ &\left|\frac{|\mathcal{I}_\ell^c|^{-1/2}\sum_{i\in\mathcal{I}_\ell^c}\left\{X_{i,j_2}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\left\{\sum_{j\neq j_2}X_{i,j}(\overline{W}^{(\ell)}_{j_1,j}-W_{0,j_1,j})\right\}}{|\mathcal{I}_\ell^c|^{-1}\sum_{i\in\mathcal{I}_\ell^c}X_{i,j_2}\left\{X_{i,j_2}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}}\right|=\frac{I_1(j_1,j_2,\ell)}{I_2(j_1,j_2,\ell)}.
\end{aligned}
$$

We next bound $I_1(j_1,j_2,\ell)$ and $I_2(j_1,j_2,\ell)$, respectively.

For $I_1(j_1,j_2,\ell)$, we have that, $I_1(j_1,j_2,\ell)\leq I_{1,1}(j_1,j_2,\ell)+I_{1,2}(j_1,j_2,\ell)$, where

$$
\begin{aligned}
I_{1,1}(j_1,j_2,\ell) &= \left||\mathcal{I}_\ell^c|^{-1/2}\sum_{i\in\mathcal{I}_\ell^c}\left\{\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\left\{\sum_{j\neq j_2}X_{i,j}(\overline{W}^{(\ell)}_{j_1,j}-W_{0,j_1,j})\right\}\right|,\\
I_{1,2}(j_1,j_2,\ell) &= \left||\mathcal{I}_\ell^c|^{-1/2}\sum_{i\in\mathcal{I}_\ell^c}\left\{X_{i,j_1}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\left\{\sum_{j\neq j_2}X_{i,j}(\overline{W}^{(\ell)}_{j_1,j}-W_{0,j_1,j})\right\}\right|.
\end{aligned}
$$

For $I_{1,1}(j_1,j_2)$, by Cauchy-Schwarz inequality, it can be upper bounded by

$$
\begin{aligned}
&|\mathcal{I}_\ell^c|^{1/2}\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)-\boldsymbol{\beta}_0^{\ell}(j_1,j_2)\|_2\|\overline{\boldsymbol{W}}^{(\ell)}_{j_1}-\boldsymbol{W}_{0,j_1}\|_2\|\boldsymbol{\Sigma}_0\|_2\\
+\ &|\mathcal{I}_\ell^c|^{1/2}\|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)-\boldsymbol{\beta}_0^{\ell}(j_1,j_2)\|_1\|\overline{\boldsymbol{W}}^{(\ell)}_{j_1}-\boldsymbol{W}_{0,j_1}\|_1\left\||\mathcal{I}_\ell^c|^{-1}\sum_{i\in\mathcal{I}_\ell^c}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}_0\right\|_{\infty,\infty},
\end{aligned}
$$

where $\|\cdot\|_{\infty,\infty}$ denotes the elementwise maximum norm in absolute value. In the proof of Lemma 4, we have shown that $\|\boldsymbol{\Sigma}_0\|_2$ is bounded. By Condition (A5), the first line is upper bounded by $O(1)\sqrt{n}n^{-(\kappa_4+\kappa_5)}$ where $O(1)$ denotes some positive constant.

In addition, using similar arguments as in Equation (A.70) of Shi et al. (2020), we have

$$
\left\||\mathcal{I}_\ell^c|^{-1}\sum_{i\in\mathcal{I}_\ell^c}\boldsymbol{X}_i\boldsymbol{X}_i^\top-\boldsymbol{\Sigma}_0\right\|_{\infty,\infty}=O(n^{-1/2}\sqrt{\log d}),
$$

with probability approaching 1. This together with the condition on the number of mediators d and Condition (A5) implies that the second line is $O(n^{-\kappa_6}\sqrt{\log n})$, with probability approaching 1. Consequently,

$$\max_{j_1,j_2,\ell} I_{1,1}(j_1,j_2,\ell) \le O(1) n^{\max(1/2-(\kappa_4+\kappa_5),-\kappa_6)}\sqrt{\log n}, \tag{S17}$$

where $O(1)$ denotes some positive constant.

For $I_{1,2}(j_1,j_2)$, note that $W_{0,j_1,j} \neq 0$ only when j is a parent of j_1. By (A4), we have $W_{0,j_1,j} \neq 0$ only when $j \in \mathrm{ACT}(j_1, \widetilde{\boldsymbol{W}}^{(\ell)})$. It follows that,

$$\sum_{j\neq j_2} X_{i,j}\left(\overline{W}^{(\ell)}_{j_1,j} - W_{0,j_1,j}\right) = \sum_{j\neq j_2, j\in \mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})} X_{i,j}\left(\overline{W}^{(\ell)}_{j_1,j} - W_{0,j_1,j}\right).$$

By the definition of $\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)$, we have $\mathrm{E}\left[\left\{X_{j_2} - \boldsymbol{X}^\top \boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\right\} X_j | \widetilde{\boldsymbol{W}}^{(\ell)}, \overline{\boldsymbol{W}}^{(\ell)}\right] = 0$ for any $j \in \mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)}) - \{j_2\}$. Note that $\widetilde{\boldsymbol{W}}^{(\ell)}$ and $\overline{\boldsymbol{W}}^{(\ell)}$ are constructed by samples in $\{X_{i,j}\}_{i\in\mathcal{I}_\ell}$. It follows that $\mathrm{E}\left\{\varphi_i^{(\ell)}(j_1,j_2)|\widetilde{\boldsymbol{W}}^{(\ell)}, \overline{\boldsymbol{W}}^{(\ell)}\right\} = 0$ for any $i \in \mathcal{I}_\ell$, where

$$\varphi_i^{(\ell)}(j_1,j_2) = \left\{X_{i,j_2} - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\right\}\left\{\sum_{j\neq j_2, j\in \mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})} X_{i,j}(\overline{W}^{(\ell)}_{j_1,j} - W_{0,j_1,j})\right\}.$$

By the definition of the Orlicz norm and Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\|\varphi_i^{(\ell)}(j_1,j_2)\|_{\psi_1|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} = \|\varphi_i^{(\ell)}(j_1,j_2)\|^2_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} \\ \le& \frac{1}{2\tau}\|X_{i,j_2} - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|^2_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} + \frac{\tau}{2}\|\boldsymbol{X}_i^\top(\overline{\boldsymbol{W}}^{(\ell)}_{j_1} - \boldsymbol{W}_{0,j_1})\|^2_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}}, \end{aligned} \tag{S18}$$

where $\|\cdot\|_{\psi_p|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}}$ denotes the Orlicz norm conditional on $\widetilde{\boldsymbol{W}}^{(\ell)}$ and $\overline{\boldsymbol{W}}^{(\ell)}$. Since j_2 does not belong to the support of $\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)$, we have by (S6) that

$$\|X_{i,j_2} - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} \le c_0\left\{1 + \|\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|_2\right\}.$$

Using similar arguments in Lemma A.1 of Shi et al. (2020), we have $\max_{j_1,j_2,\ell}\|\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|_2 = O(1)$. Henceforth, $\max_{i,j_1,j_2,\ell}\|X_{i,j_2} - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} = O(1)$. Similarly, we have,

$$\max_{i,j_1,j_2,\ell}\|\boldsymbol{X}_i^\top(\overline{\boldsymbol{W}}^{(\ell)}_{j_1} - \boldsymbol{W}_{0,j_1})\|_{\psi_2|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} \le c_0 \max_{j_1}\|\overline{\boldsymbol{W}}^{(\ell)}_{j_1} - \boldsymbol{W}_{0,j_1}\|_2 \le c_0\kappa_3 n^{-\kappa_5},$$

by Condition (A5). Setting $\tau = n^{-\kappa_5}$, it follows from (S18) that

$$\max_{i,\ell}\|\varphi_i^{(\ell)}(j_1,j_2)\|_{\psi_1|\widetilde{\boldsymbol{W}}^{(\ell)},\overline{\boldsymbol{W}}^{(\ell)}} = O(n^{-\kappa_5}), \tag{S19}$$

by (A4) and (A5). It follows from Lemma G.3 of Shi et al. (2018) that

$$\Pr\left(\left|\sum_{i\in\mathcal{I}_\ell^c}\varphi_i^{(\ell)}(j_1,j_2)\right| > t \,|\, \max_j \|\overline{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_2 \le \kappa_3 n^{-\kappa_5}, \widetilde{\boldsymbol{W}}^{(\ell)}\right) \le 2\exp\left\{-\kappa\min\left(\frac{t^2}{n^{1-2\kappa_5}}, \frac{t}{n^{-\kappa_5}}\right)\right\},$$

for some constant $\kappa > 0$. By Bonferroni's inequality, we have that,

$$\begin{aligned}
&\Pr\left(\max_{j_1,j_2,\ell}\left|\sum_{i\in\mathcal{I}_\ell^c}\varphi_i^{(\ell)}(j_1,j_2)\right| > t \,|\, \max_j \|\overline{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_2 \le \kappa_3 n^{-\kappa_5}, \widetilde{\boldsymbol{W}}^{(\ell)}\right) \\
\le\;& 4(d+2)^2\exp\left\{-c\min\left(\frac{t^2}{n^{1-2\kappa_5}}, \frac{t}{n^{-\kappa_5}}\right)\right\} \le 4\exp\left\{-c\min\left(\frac{t^2}{n^{1-2\kappa_5}}, \frac{t}{n^{-\kappa_5}}\right) + 2\log(d+2)\right\}.
\end{aligned}$$

Setting $t = 3\kappa_1 c^{-1} n^{1/2-\kappa_5}\sqrt{\log n}$, for a sufficiently large n, we have that,

$$\begin{aligned}
4\exp\left\{-c\min\left(\frac{t^2}{n^{1-2\kappa_5}}, \frac{t}{n^{-\kappa_5}}\right) + 2\log(d+2)\right\} &= 4\exp\{-3\kappa_1\log n + 2\log(d+2)\} \\
&= \frac{4(d+2)^2}{n^{3\kappa_1}} = O(n^{-\kappa_1}) = o(1).
\end{aligned}$$

Therefore, conditional on the events in (A4) and (A5), we have that, with probability approaching one, $\max_{j_1,j_2,\ell}|\sum_{i\in\mathcal{I}_\ell^c}\varphi_i^{(\ell)}(j_1,j_2)| = O(n^{1/2-\kappa_5}\sqrt{\log n})$, or equivalently,

$$\max_{j_1,j_2,\ell} I_{1,2}(j_1,j_2,\ell) = O(n^{-\kappa_5}\sqrt{\log n}). \tag{S20}$$

Combining (S17) and (S20) together yields, for some constant $\kappa > 0$,

$$\max_{j_1,j_2,\ell} I_1(j_1,j_2,\ell) = O(n^{-\kappa}), \tag{S21}$$

For $I_2(j_1,j_2,\ell)$, we first define

$$\begin{aligned}
I_2^*(j_1,j_2,\ell) &= |\mathcal{I}_\ell^c|^{-1}\sum_{i\in\mathcal{I}_\ell^c}\left\{X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}X_{i,j_2}, \\
I_2^{**}(j_1,j_2,\ell) &= \mathrm{E}\left\{I_2^*(j_1,j_2,\ell)|\widetilde{\boldsymbol{W}}^{(\ell)}\right\} = \mathrm{E}\left[\left\{X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}^2 \mid \widetilde{\boldsymbol{W}}^{(\ell)}\right].
\end{aligned}$$

Following the proof of Corollaries 4.1 and 4.2 of Ning and Liu (2017), we have that,

$$\begin{aligned}
&|I_2(j_1,j_2,\ell) - I_2^*(j_1,j_2,\ell)| \le |\mathcal{I}_\ell^c|^{-1}\left|\sum_{i\in\mathcal{I}_\ell^c}X_{i,j_2}\{\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2) - \widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\}^\top\boldsymbol{X}_i\right| \\
\le\;& \|\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2) - \widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\|_1\left(\frac{2}{n}\left\|\sum_{i\in\mathcal{I}_\ell^c}X_{i,j_2}\boldsymbol{X}_i - |\mathcal{I}_\ell^c|\mathrm{E}X_{j_2}\boldsymbol{X}\right\|_\infty\right) \\
+\;& \|\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2) - \widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\|_2\lambda_{\max}(\boldsymbol{\Sigma}_0) = O(n^{-\kappa_4}) + O(n^{-\kappa_6-1/2}\sqrt{\log n}),
\end{aligned} \tag{S22}$$

where the big-O term is uniform in (j_1, j_2, ℓ). In addition, note that

$$\mathrm{E}|\mathcal{I}_\ell^c|^{-1} \sum_{i \in \mathcal{I}_\ell^c} \left[X_{i,j_2} \left\{ X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X}_i \right\} | \widetilde{\boldsymbol{W}}^{(\ell)} \right] = \mathrm{E}\left[X_{j_2} \left\{ X_{j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X} \right\} | \widetilde{\boldsymbol{W}}^{(\ell)} \right]$$

$$= \mathrm{E}\left[\left\{ X_{j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X} \right\}^2 | \widetilde{\boldsymbol{W}}^{(\ell)} \right] \geq \|\{1, \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\}\|_2^2 \lambda_{\min}(\Sigma_0) \geq \lambda_{\min}(\Sigma_0).$$

Since the minimum eigenvalue of Σ_0 is bounded away from zero, we have

$$\min_{j_1, j_2, \ell} I_2^{**}(j_1, j_2, \ell) = \min_{j_1, j_2, \ell} \mathrm{E}\left\{ I_2^*(j_1, j_2, \ell) | \widetilde{\boldsymbol{W}}^{(\ell)} \right\} \geq 2\varepsilon, \tag{S23}$$

for some $\varepsilon > 0$. Similar to (S20), we have,

$$\max_{_1, j_2, \ell} \left\| \sum_{i \in \mathcal{I}_\ell^c} \{X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X}_i\} X_{i,j_2} - |\mathcal{I}_\ell^c| I_2^{**}(j_1, j_2, \ell) \right\|_\infty = O(\sqrt{n \log n}). \tag{S24}$$

This together with (S23) yields,

$$\min_{j_1, j_2, \ell} I_2^*(j_1, j_2, \ell) = \min_{j_1, j_2, \ell} |\mathcal{I}_\ell^c|^{-1} \sum_{i \in \mathcal{I}_\ell^c} \{X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X}_i\} X_{i,j_2} \geq 2\varepsilon.$$

Combining this together with (S22), we have,

$$\begin{aligned} \min_{j_1, j_2, \ell} I_2(j_1, j_2, \ell) \geq \min_{j_1, j_2, \ell} \Bigg[& |\mathcal{I}_\ell^c|^{-1} \sum_{i \in \mathcal{I}_\ell^c} \left\{ X_{i,j_2} - \boldsymbol{\beta}_0^{(\ell)\top}(j_1, j_2)\boldsymbol{X}_i \right\} X_{i,j_2} \\ & + |\mathcal{I}_\ell^c|^{-1} \sum_{i \in \mathcal{I}_\ell^c} X_{i,j_2} \left\{ \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2) - \widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) \right\}^\top \boldsymbol{X}_i \Bigg] \geq \varepsilon. \end{aligned} \tag{S25}$$

Combining (S21) and (S25) together yields,

$$\max_{(j_1, j_2) \in \mathcal{S}(q_1, q_2, \widehat{\boldsymbol{B}}^{(\ell)})} |\sqrt{|\mathcal{I}_\ell^c|}(\widehat{W}_{j_1, j_2} - W_{0, j_1, j_2}) - \eta_{j_1, j_2}^{(\ell)}| \leq O(1) n^{-\kappa}, \tag{S26}$$

for some constant $\kappa > 0$. This gives the first desired upper bound.

To obtain the second upper bound, we have that,

$$\begin{aligned} \max_{(j_1, j_2) \in \mathcal{S}(q_1, q_2, \widehat{\boldsymbol{B}}^{(\ell)})} |\eta_{j_1, j_2}^{(\ell)} - \eta_{0, j_1, j_2}^{(\ell)}| \leq \max_{(j_1, j_2) \in \mathcal{S}(q_1, q_2, \widehat{\boldsymbol{B}}^{(\ell)})} \frac{|\mathcal{I}_\ell^c|^{-1/2} |\sum_{i \in \mathcal{I}_\ell^c} \{\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\}^\top \boldsymbol{X}_i \varepsilon_{i,j_1}|}{I_2(j_1, j_2, \ell)} \\ + \max_{(j_1, j_2) \in \mathcal{S}(q_1, q_2, \widehat{\boldsymbol{B}}^{(\ell)})} \frac{|I_2(j_1, j_2, \ell) - I_2^{**}(j_1, j_2, \ell)| |\mathcal{I}_\ell^c|^{-1/2} |\sum_{i \in \mathcal{I}_\ell^c} \{\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1, j_2) - \boldsymbol{\beta}_0^{(\ell)}(j_1, j_2)\}^\top \boldsymbol{X}_i \varepsilon_{i,j_1}|}{I_2(j_1, j_2, \ell) I_2^{**}(j_1, j_2, \ell)}. \end{aligned}$$

By (S22), (S23), (S24) and (S25), we have,

$$\begin{aligned}
&\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} |\eta^{(\ell)}_{j_1,j_2} - \eta^{(\ell)}_{0,j_1,j_2}| \\
&\le \frac{2}{\varepsilon} \max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} \left| |\mathcal{I}^c_\ell|^{-1/2} \sum_{i\in\mathcal{I}^c_\ell} \{\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2) - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\}^\top \boldsymbol{X}_i \varepsilon_{i,j_1} \right|.
\end{aligned} \tag{S27}$$

Note that the supports of $\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)$ and $\boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)$ belong to $\mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})$. Since ε_{i,j_1} is independent of $\{X_{i,j} : j\in \mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})\}$ conditional on $\widetilde{\boldsymbol{W}}^{(\ell)}$, we have,

$$\begin{aligned}
&\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} \left| |\mathcal{I}^c_\ell|^{-1/2} \sum_{i\in\mathcal{I}^c_\ell} \{\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2) - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\}^\top \boldsymbol{X}_i \varepsilon_{i,j_1} \right| \\
\le\ & \max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} \|\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2) - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\|_1 \max_{j\in\mathrm{ACT}(j_1,\widehat{\boldsymbol{B}}^{(\ell)})} \left| |\mathcal{I}^c_\ell|^{-1/2} \sum_{i\in\mathcal{I}^c_\ell} X_{i,j}\varepsilon_{i,j_1} \right|.
\end{aligned}$$

By (A4), ε_{i,j_1} is conditionally independent of X_j for any $j\in\mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})$. Similar to (S20), by Bernstein's inequality, we have that,

$$\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} \max_{j\in\mathrm{ACT}(j_1,\widetilde{\boldsymbol{W}}^{(\ell)})} \left| |\mathcal{I}^c_\ell|^{-1/2} \sum_{i\in\mathcal{I}^c_\ell} X_{i,j}\varepsilon_{i,j_1} \right| = O(\sqrt{\log n}).$$

It follows from Condition (A5) that,

$$\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} \left| |\mathcal{I}^c_\ell|^{-1/2} \sum_{i\in\mathcal{I}^c_\ell} \{\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2) - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\}^\top \boldsymbol{X}_i \varepsilon_{i,j_1} \right| = O(n^{-\kappa}),$$

for some constant $\kappa>0$. In view of (S27), we obtain the second desired upper bound,

$$\max_{(j_1,j_2)\in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})} |\eta^{(\ell)}_{j_1,j_2} - \eta^{(\ell)}_{0,j_1,j_2}| = O(n^{-\kappa}), \tag{S28}$$

for some constant $\kappa>0$.

Finally, combining (S26) and (S28) together yields (S10). This completes Step 1.

Step 2: Recall that, each ε_{i,j_1} is uncorrelated with $X_{i,j_2} - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\boldsymbol{X}_i$. Since both ε_{i,j_1} and $X_{i,j_2} - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\boldsymbol{X}_i$ are normally distributed given $\widetilde{\boldsymbol{W}}^{(\ell)}$, they are independent as well. As a result, we have,

$$\mathrm{Var}\left[\left\{X_{i,j_2} - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\boldsymbol{X}_i\right\}\varepsilon_{i,j_1} | \widetilde{\boldsymbol{W}}^{(\ell)}\right] = \sigma^2_* \mathrm{E}\left[\left\{X_{i,j_2} - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\boldsymbol{X}_i\right\}^2 \mid \widetilde{\boldsymbol{W}}^{(\ell)}\right],$$

for $1\le j_1\le d+1$. Recall that $I^{**}_2(j_1,j_2,\ell) = \mathrm{E}\left[\left\{X_{i,j_2} - \boldsymbol{\beta}^{(\ell)\top}_0(j_1,j_2)\boldsymbol{X}_i\right\}^2 \mid \widetilde{\boldsymbol{W}}^{(\ell)}\right]$. Then, we have for $1\le j_1\le d+1$ that

$$\begin{aligned}\operatorname{Var}\left(\eta_{0,j_1,j_2}^{(\ell)}|\widetilde{\boldsymbol{W}}^{(\ell)}\right) &= \frac{1}{\{I_2^{**}(j_1,j_2,\ell)\}^2}\operatorname{Var}\left[\left\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\varepsilon_{i,j_1}\mid\widetilde{\boldsymbol{W}}^{(\ell)}\right] \\ &= \frac{1}{\sigma_*^2\mathrm{E}\left[\left\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}^2\mid\widetilde{\boldsymbol{W}}^{(\ell)}\right]}.\end{aligned} \tag{S29}$$

Since $j_2 \notin \mathcal{M}^{(\ell)}(j_1,j_2)$, we have

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_0)\left\|\{1,\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\}^\top\right\|_2^2 &\leq \mathrm{E}\left[\left\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}^2\mid\widetilde{\boldsymbol{W}}^{(\ell)}\right] \\ &\leq \lambda_{\max}(\boldsymbol{\Sigma}_0)\left\|(1,\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2))^\top\right\|_2^2.\end{aligned}$$

Following the proof of Lemma A.1 of Shi et al. (2020), we have $\max_{j_1,j_2,\ell}\|\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)\|_2 = O(1)$. Since $\lambda_{\min}(\boldsymbol{\Sigma}_0)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_0)$ are uniformly bounded away from zero and infinity, there exists some constant $\kappa \geq 1$ such that, for any j_1, j_2,

$$\kappa^{-1} \leq \mathrm{E}[\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\}^2|\widetilde{\boldsymbol{W}}^{(\ell)}] \leq \kappa. \tag{S30}$$

It follows from (S29) that, for any j_1, j_2,

$$\kappa^{-1} \leq \operatorname{Var}(\eta_{0,j_1,j_2}^{(\ell)}|\widetilde{\boldsymbol{W}}^{(\ell)}) \leq \kappa. \tag{S31}$$

We index all pairs of indices (j_1,j_2) in $\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})$ by $\{j_1(0),j_2(0)\},\{j_1(1),j_2(1)\},\ldots,\{j_1(L-1),j_2(L-1)\}$, where $L = |\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})|$. Next, define a covariance matrix $V_0 \in \mathbb{R}^{L\times L}$, such that its (l_1,l_2)th entry is the covariance of $\eta_{0,j_1(l_1),j_2(l_1)}^{(\ell)}$ and $\eta_{0,j_1(l_2),j_2(l_2)}^{(\ell)}$ conditional on $\widetilde{\boldsymbol{W}}^{(\ell)}$. By (S31), the diagonal elements of V_0 are uniformly bounded away from zero and infinity. We also comment that $\eta_{j_1,j_2}^{(\ell)}$ defined in (15) in Section 4.4 of the paper can be viewed as a more intuitive version of $\eta_{0,j_1,j_2}^{(\ell)}$ defined in (S11).

When L is finite, by the classical Lindeberg-Feller central limit theorem and Condition (A4), we have (S13) holds. When L diverges, we have $L \leq d^2$ that grows at an polynomial order of n. By (S31), we have $\min_{j_1,j_2,\ell}\operatorname{Var}\left(\eta_{0,j_1,j_2}^{(\ell)}|\widetilde{\boldsymbol{W}}^{(\ell)}\right)$ is uniformly bounded away from zero. Moreover, following similar arguments in proving (S19), we have $\max_{j_1,j_2,\ell}\|\eta_{0,j_1,j_2}^{(\ell)}\|_{\psi_1|\widetilde{\boldsymbol{W}}^{(\ell)}} = O(1)$. Then, by Corollary 2.1 of Chernozhukov et al. (2013), (S13) holds as well. This completes Step 2.

Step 3: Define a covariance matrix $\widehat{V} \in \mathbb{R}^{L\times L}$, such that its (l_1,l_2)th entry is the covariance of $\eta_{j_1(l_1),j_2(l_1)}^{*(\ell)}$ and $\eta_{j_1(l_2),j_2(l_2)}^{*(\ell)}$ conditional on the data. To bound $\|\widehat{V}-V_0\|_{\infty,\infty}$ in (S14), it suffices to bound

$$\max_{\substack{(j_1,j_2),(j_3,j_4)\\ \in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}} \left|\mathrm{cov}\left[\eta^{*(\ell)}_{j_1,j_2},\eta^{*(\ell)}_{j_3,j_4}|\{X_{i,j}\}_{1\le i\le n,0\le j\le d+1}\right]-\mathrm{cov}\left[\eta^{(\ell)}_{0,j_1,j_2},\eta^{(\ell)}_{0,j_3,j_4}|\{X_{i,j}\}_{i\in\mathcal{I}_\ell^c,0\le j\le d+1}\right]\right|.$$

Recall in Section 4.4, when $j_1\neq j_3$, we have $\mathrm{cov}(\eta^{*(\ell)}_{j_1,j_2},\eta^{*(\ell)}_{j_3,j_4}|\{X_{i,j}\}_{1\le i\le n,0\le j\le d+1})=0$. Similarly, $\mathrm{cov}(\eta^{(\ell)}_{0,j_1,j_2},\eta^{(\ell)}_{0,j_3,j_4}|\{X_{i,j}\}_{i\in\mathcal{I}_\ell^c,0\le j\le d+1})=0$ when $j_1\neq j_3$. As a result, we have

$$\begin{aligned}\|\widehat{V}-V_0\|_{\infty,\infty} = \max_{\substack{(j_1,j_2),(j_1,j_3)\\ \in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}} &\left|\mathrm{cov}\left[\eta^{*(\ell)}_{j_1,j_2},\eta^{*(\ell)}_{j_1,j_3}|\{X_{i,j}\}_{1\le i\le n,0\le j\le d+1}\right]\right.\\ &\left.-\mathrm{cov}\left[\eta^{(\ell)}_{0,j_1,j_2},\eta^{(\ell)}_{0,j_1,j_3}|\{X_{i,j}\}_{i\in\mathcal{I}_\ell^c,0\le j\le d+1}\right]\right|.\end{aligned}$$

After some calculations, we have, for $1\le j_1\le d+1$,

$$\begin{aligned}\mathrm{cov}\left[\eta^{(\ell)}_{0,j_1,j_2},\eta^{(\ell)}_{0,j_1,j_3}|\{X_{i,j}\}_{i\in\mathcal{I}_\ell^c,0\le j\le d+1}\right] &= \frac{\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}|\widetilde{\boldsymbol{W}}^{(\ell)}\right]\sigma_*^2}{I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)},\\ \mathrm{cov}\left[\eta^{*(\ell)}_{j_1,j_2},\eta^{*(\ell)}_{j_1,j_3}|\{X_{i,j}\}_{1\le i\le n,0\le j\le d+1}\right] &= \frac{|\mathcal{I}_\ell^c|^{-1}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\widehat{\sigma}_*^2}{I_2(j_1,j_2,\ell)I_2(j_1,j_3,\ell)}.\end{aligned}$$

It follows that

$$\begin{aligned}\|\widehat{V}-V_0\|_{\infty,\infty} &= \max_{\substack{(j_1,j_2),(j_1,j_3)\\ \in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}} I_{3,1}(j_1,j_2,j_3,\ell)+\max_{\substack{(j_1,j_2),(j_1,j_3)\\ \in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}} I_{3,2}(j_1,j_2,j_3,\ell)\\ &\quad+\max_{\substack{(j_1,j_2),(j_1,j_3)\\ \in\mathcal{S}(q_1,q_2,\widehat{\boldsymbol{B}}^{(\ell)})}} I_{3,3}(j_1,j_2,j_3,\ell),\end{aligned}$$

where

$$\begin{aligned}I_{3,1}(j_1,j_2,j_3,\ell) &= \frac{\left|\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\ |\ \widetilde{\boldsymbol{W}}^{(\ell)}\right]\right|}{I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)}|\widehat{\sigma}_*^2-\sigma_*^2|,\\ I_{3,2}(j_1,j_2,j_3,\ell) &= \frac{\widehat{\sigma}_*^2}{I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)}\left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\right.\\ &\quad\left.-\mathrm{E}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\right]\right|,\\ I_{3,3}(j_1,j_2,j_3,\ell) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\widehat{\sigma}_*^2\right|\\ &\quad\times\frac{|I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)-I_2(j_1,j_2,\ell)I_2(j_1,j_3,\ell)|}{I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)I_2(j_1,j_2,\ell)I_2(j_1,j_3,\ell)}.\end{aligned}$$

We next bound $\max_{j_1,j_2,j_3,\ell} I_{3,q}(j_1,j_2,j_3,\ell)$, $q=1,2,3$, respectively.

For $\max_{j_1,j_2,j_3,\ell} I_{3,1}(j_1,j_2,j_3,\ell)$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\max_{j_1,j_2,j_3,\ell}\left|\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\mid \widetilde{\boldsymbol{W}}^{(\ell)}\right]\right|\\ &\le \max_{j_1,j_2,j_3,\ell}\prod_{k\in\{j_2,j_3\}}\mathrm{E}\left[\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}^2\mid \widetilde{\boldsymbol{W}}^{(\ell)}\right].\end{aligned}$$

By (S30), we have,

$$\max_{j_1,j_2,j_3,\ell}\left|\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\mid \widetilde{\boldsymbol{W}}^{(\ell)}\right]\right|=O(1). \tag{S32}$$

Combining (S32) with (S23) and Lemma 4 yields,

$$\max_{j_1,j_2,j_3,\ell} I_{3,1}(j_1,j_2,j_3,\ell)=O(n^{-\kappa_7}). \tag{S33}$$

For $\max_{j_1,j_2,j_3,\ell} I_{3,2}(j_1,j_2,j_3,\ell)$, by Lemma 4, we have $\widehat{\sigma}_*^2=O(1)$. Define

$$\begin{aligned}I_3^*(j_1,j_2,j_3,\ell) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\right.\\ &\quad\left.-\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\mid \widetilde{\boldsymbol{W}}^{(\ell)}\right]\right|.\\ I_3^{**}(j_1,j_2,j_3) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right.\right.\\ &\quad\left.\left.-\prod_{k\in\{j_2,j_3\}}\left\{X_{i,k}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\right|.\end{aligned}$$

Similar to (S20), we have,

$$\begin{aligned}\max_{j_1,j_2,j_3,\ell}\left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\{X_{i,k}-\boldsymbol{\beta}_0^\top(j_1,k)\boldsymbol{X}_i\}-\mathrm{E}\left[\prod_{k\in\{j_2,j_3\}}\left\{X_k-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,k)\boldsymbol{X}_0\right\}\mid \widetilde{\boldsymbol{W}}^{(\ell)}\right]\right|\\ =O(n^{-1/2}\sqrt{\log n}),\end{aligned}$$

Next, $\max_{j_1,j_2,j_3,\ell} I_3^{**}(j_1,j_2,j_3,\ell)\le I_{4,1}(j_1,j_2,j_3,\ell)+I_{4,2}(j_1,j_2,j_3,\ell)+I_{4,3}(j_1,j_2,j_3,\ell)$, where

$$
\begin{aligned}
I_{4,1}(j_1,j_2,j_3,\ell) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\left\{\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)-\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\right\}^\top \boldsymbol{X}_i\left\{X_{i,j_3}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_3)\boldsymbol{X}_i\right\}\right|,\\
I_{4,2}(j_1,j_2,j_3,\ell) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\left\{\boldsymbol{\beta}_0^{(\ell)}(j_1,j_3)-\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_3)\right\}^\top \boldsymbol{X}_i\left\{X_{i,j_2}-\boldsymbol{\beta}_0^{(\ell)\top}(j_1,j_2)\boldsymbol{X}_i\right\}\right|,\\
I_{4,3}(j_1,j_2,j_3,\ell) &= \left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\left\{\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)-\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\right\}^\top \boldsymbol{X}_i\boldsymbol{X}_i^\top\left\{\boldsymbol{\beta}_0^{(\ell)}(j_1,j_2)-\widehat{\boldsymbol{\beta}}^{(\ell)}(j_1,j_2)\right\}\right|.
\end{aligned}
$$

Using similar arguments in proving (S17) and (S22), we have

$$
\begin{aligned}
\max_{j_1,j_2,j_3,\ell} I_{4,1}(j_1,j_2,j_3,\ell) &= O(n^{-\kappa_5}),\\
\max_{j_1,j_2,j_3,\ell} I_{4,2}(j_1,j_2,j_3,\ell) &= O(n^{-\kappa_5}),\\
\max_{j_1,j_2,j_3,\ell} I_{4,3}(j_1,j_2,j_3,\ell) &= O(n^{-2\kappa_5}).
\end{aligned}
$$

Therefore, we have, for some constant $\kappa>0$,

$$
\max_{j_1,j_2,j_3,\ell} I_2^*(j_1,j_2,j_3,\ell) = O(n^{-\kappa}). \tag{S34}
$$

It then follows from (S23) that, for some constant $\kappa>0$,

$$
\max_{j_1,j_2,j_3,\ell} I_2(j_1,j_2,j_3,\ell) = O(n^{-\kappa}). \tag{S35}
$$

For $\max_{j_1,j_2,j_3,\ell} I_{3,3}(j_1,j_2,j_3,\ell)$, combining (S34) with (S32), we have,

$$
\left|\frac{1}{|\mathcal{I}_\ell^c|}\sum_{i\in\mathcal{I}_\ell^c}\prod_{k\in\{j_2,j_3\}}\left[\left\{X_{i,k}-\widehat{\boldsymbol{\beta}}^{(\ell)\top}(j_1,k)\boldsymbol{X}_i\right\}\right]\right| = O(1). \tag{S36}
$$

Using similar arguments in bounding $\max_{j_1,j_2,\ell}|I_2(j_1,j_2,\ell)-I_2^{**}(j_1,j_2,\ell)|$ in Step 2, we get

$$
\max_{j_1,j_2,j_3,\ell}|I_2^{**}(j_1,j_2,\ell)I_2^{**}(j_1,j_3,\ell)-I_2(j_1,j_2,\ell)I_2(j_1,j_3,\ell)| = O(n^{-\kappa}),
$$

for some constant $\kappa>0$. Combining this together with (S36), (S23), (S25) and that $\widehat{\sigma}_*^2=O(1)$, we have, for some constant $\kappa>0$,

$$
\max_{j_1,j_2,j_3,\ell} I_{3,3}(j_1,j_2,j_3,\ell) = O(n^{-\kappa}). \tag{S37}
$$

Combining (S33), (S35) and (S37) together yields the bound for $\|\widehat{V}-V_0\|_{\infty,\infty}$ in (S14). This completes Step 3. $\square$

S7 Proof of Theorem 2

Recall the proposed bootstrap procedure repeatedly generate random variables from $N(0,\widehat{V})$, and the critical value $\widehat{c}^{(\ell)}(q_1,q_2)$ is the upper $(\alpha/2)$th quantile of $\|N(0,\widehat{V})\|_\infty$. That is,

$$\Pr\left(\|N(0,\widehat{V})\|_\infty \le \widehat{c}^{(\ell)}(q_1,q_2)|\widehat{V}\right) = \frac{\alpha}{2}. \tag{S38}$$

We have shown in the proof of Theorem 1 that the diagonal elements of $\widehat{V}$ are uniformly bounded by some constant $\kappa > 0$. It follows from Bonferroni's inequality that

$$\begin{aligned}\Pr\left(\|N(0,\widehat{V})\|_\infty > t\sqrt{\kappa \log n}\right) &\le (d+2) \max_{j\in\{0,\ldots,d+1\}} \Pr\left(|N(0,\widehat{V}_{j,j})| > t\sqrt{\kappa\log n}\right) \\ &\le (d+2)\left\{1-\Phi\left(t\sqrt{\log n}\right)\right\},\end{aligned}$$

where $\widehat{V}_{j,j}$ is the (j,j)th entry of $\widehat{V}$. For $t\ge 1$ and $n\ge 3$, we have $t\sqrt{\log n}\ge 1$, and hence

$$\begin{aligned}1-\Phi\left(t\sqrt{\log n}\right) &= \frac{1}{\sqrt{2\pi}}\int_{t\sqrt{\log n}}^{+\infty} \exp\left(-\frac{x^2}{2}\right)dx \le \int_{t\sqrt{\log n}}^{+\infty} x\exp\left(-\frac{x^2}{2}\right)dx \\ &= \exp\left(-\frac{t^2\log n}{2}\right) = n^{-t^2/2}.\end{aligned}$$

Setting $t = 2\sqrt{\kappa_1+1}$, it follows from the condition $d = O(n^{\kappa_1})$ that,

$$\Pr\left(\|N(0,\widehat{V})\|_\infty > \sqrt{2(\kappa_1+1)c\log n}\right) = O(d/n^{\kappa_1+1}) = o(1).$$

In view of (S38), we obtain $\widehat{c}^{(\ell)}(q_1,q_2) \le \sqrt{2(\kappa_1+1)c\log n}$.

According to our test procedure, we reject the null if $\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0} > \widehat{c}(0,q)$ and $\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{d+1,q} > \widehat{c}(q,d+1)$, for some $\ell = 1,2$, where $\widehat{\boldsymbol{W}}^{*(\ell)} = |\widehat{\boldsymbol{W}}^{(\ell)}| \oplus |\widehat{\boldsymbol{W}}^{(\ell)}|^{(2)} \oplus \ldots \oplus |\widehat{\boldsymbol{W}}^{(\ell)}|^{(d)}$. To prove Theorem 2, it suffices to show

$$\begin{aligned}\Pr(\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0} \ge \sqrt{2(\kappa_1+1)c\log n}) \to 1, \\ \Pr(\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{d+1,q} \ge \sqrt{2(\kappa_1+1)c\log n}) \to 1,\end{aligned} \tag{S39}$$

under the alternative hypothesis H_1. We next prove (S39).

Under the given conditions of the theorem, there exists a path, $X_0 \to X_{i_1} \to \ldots \to X_{i_k} \to X_q$, such that $|W_{0,0,i_1}|, |W_{0,i_1,i_2}|, \ldots, |W_{0,i_{k-1},i_k}|, |W_{0,i_k,q}| \gg n^{-1/2}\sqrt{\log n}$. Under the given conditions, we have with probability tending to 1 that,

$$\max\left(|\widehat{W}^{(\ell)}_{0,i_1} - W_{0,0,i_1}|, \max_{1\le j\le k-1}|\widehat{W}^{(\ell)}_{i_j,i_{j+1}} - W_{0,i_j,i_{j+1}}|, |\widehat{W}^{(\ell)}_{i_k,q} - W_{0,i_k,q}|\right) = O(n^{-1/2}\sqrt{\log n}).$$

It follows that

$$\min\left(|\widehat{W}^{(\ell)}_{0,i_1}|, \min_{1\le j\le k-1}|\widehat{W}^{(\ell)}_{i_j,i_{j+1}}|, |\widehat{W}^{(\ell)}_{i_k,q}|\right) \gg n^{-1/2}\sqrt{\log n}. \tag{S40}$$

By definition, we have that,

$$\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0} \ge \sqrt{n}\min\left(|\widehat{W}^{(\ell)}_{0,i_1}|, \min_{1\le j\le k-1}|\widehat{W}^{(\ell)}_{i_j,i_{j+1}}|, |\widehat{W}^{(\ell)}_{i_k,q}|\right).$$

It follows from (S40) that $\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0} \gg \sqrt{\log n}$ under Condition (A5). Therefore, $\Pr\left(\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0} \ge \sqrt{2(\kappa_1+1)c\log n}\right) \to 1$, which proves the first result in (S39). Similarly, we can prove the second result in (S39) that $\Pr\left(\sqrt{n}(\widehat{\boldsymbol{W}}^{*(\ell)})_{d+1,q} \ge \sqrt{2(\kappa_1+1)c\log n}\right)$ $\to 1$. This completes the proof of Theorem 2. $\square$

S8 Proof of Theorem 3

First, we show that the p-values $\widehat{p}^{(\ell)}(0,q)$ and $\widehat{p}^{(\ell)}(q,d+1)$ are asymptotically independent. Let $\mathrm{ACT}(q)$ and $\mathrm{DES}(q)$ denote the set of ancestors and descendants of X_q, respectively. By Condition (A4), the two test statistics $\sqrt{\mathcal{I}_\ell^c}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0}$ and $\sqrt{\mathcal{I}_\ell^c}(\widehat{\boldsymbol{W}}^{*(\ell)})_{d+1,q}$ are constructed based on the set of estimators $\left\{\widehat{W}^{(\ell)}_{j_1,j_2} : j_1 \in \mathrm{ACT}(q)\cup\{q\}, j_2 \in \mathrm{ACT}(q)\right\}$ and $\left\{\widehat{W}^{(\ell)}_{j_1,j_2} : j_1 \in \mathrm{DES}(q), j_2 \in \mathrm{DES}(q)\cup\{q\}\right\}$, respectively. These two sets of estimators are asymptotically independent conditional on $\{X_i\}_{i\in\mathcal{I}_\ell}$. To better illustrate this, note that for any (j_1,j_2,j_3,j_4) such that $j_1 \in \mathrm{ACT}(q)\cup\{q\}$, $j_2 \in \mathrm{ACT}(q)$, $j_3 \in \mathrm{DES}(q)$, $j_4 \in \mathrm{DES}(q)\cup\{q\}$, we have $j_1 \neq j_3$. As discussed in Section 4.4, $\sqrt{n}(\widehat{W}^{(\ell)}_{j_1,j_2} - W_{0,j_1,j_2})$ and $\sqrt{n}(\widehat{W}^{(\ell)}_{j_3,j_4} - W_{0,j_3,j_4})$ are asymptotically uncorrelated. Since the two variables are jointly normal, they are asymptotically independent as well. As a result, the two test statistics $\sqrt{\mathcal{I}_\ell^c}(\widehat{\boldsymbol{W}}^{*(\ell)})_{q,0}$ and $\sqrt{\mathcal{I}_\ell^c}(\widehat{\boldsymbol{W}}^{*(\ell)})_{d+1,q}$ are asymptotically independent, and so are their corresponding p-values.

Next, a closer look at the proof of Theorem 1 shows that the type-I error rates can be uniformly controlled across different mediators. That is,

$$\max_{\ell\in\{1,2\}}\max_{q\in\{1,\dots,d\}}\Pr\Big\{\widehat{p}^{(\ell)}_{\max}(q) \le \alpha \mid H_0(q) \text{ holds}\Big\} \le \alpha + o(1), \tag{S41}$$

for any given significance level $0<\alpha<1$. Following the proof of Theorem 1 in Djordjilović et al. (2019), and by (S41), we can show that,

$$\max_{\ell\in\{1,2\}}\max_{q\in\{1,\dots,d\}}\Pr\Big\{\widehat{p}^{(\ell)}_{\max}(q) \le \alpha \mid H_0(q) \text{ holds},\ \ \widehat{p}^{(\ell)}_{\min}(q) \le c^{(\ell)}\Big\} \le \alpha + o(1), \tag{S42}$$

for any significance level α and the critical value $c^{(\ell)} > 0$.

Then, similar to Theorem 1.3 of Benjamini and Yekutieli (2001), and by (S42), we have that $\mathrm{FDR}(\mathcal{H}^{(\ell)})$ is guaranteed at level $\alpha/2$. Recall $\mathcal{H} = \mathcal{H}^{(1)} \cup \mathcal{H}^{(2)}$ is the final set of our selected mediators, and $\mathcal{N}$ is the set of unimportant mediators. It follows that,

$$\begin{aligned}\mathrm{FDR}(\mathcal{H}) &= \mathrm{E}\left(\frac{|\mathcal{N} \cup \mathcal{H}|}{\max(1, \mathcal{H})}\right) \leq \sum_{\ell \in \{1,2\}} \mathrm{E}\left(\frac{|\mathcal{N} \cup \mathcal{H}^{(\ell)}|}{\max(1, \mathcal{H})}\right) \leq \sum_{\ell \in \{1,2\}} \mathrm{E}\left(\frac{|\mathcal{N} \cup \mathcal{H}^{(\ell)}|}{\max(1, \mathcal{H}^{(\ell)})}\right) \\ &= \sum_{\ell \in \{1,2\}} \mathrm{FDR}(\mathcal{H}^{(\ell)}) = \alpha + o(1).\end{aligned}$$

This completes the proof of Theorem 3. □

S9 Proof of Proposition 1

Outline of the proof: We choose $\lambda = \kappa\sqrt{\log n/n}$ for some sufficiently large constant $\kappa > 0$. We divide the proof into three steps. In Step 1, we establish the uniform convergence rate of $\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi) - \boldsymbol{W}_j(\pi)\|_2$. Due to the acyclic constraint, $\widetilde{\boldsymbol{W}}^{(\ell)} = \widetilde{\boldsymbol{W}}^{(\ell)}(\pi)$ for some ordering π. In Step 2, we show that

$$\begin{aligned}&\min_{\pi \notin \Pi_0}\left(\sum_{i \in \mathcal{I}_\ell} \|\boldsymbol{X}_i - \widetilde{\boldsymbol{W}}^{(\ell)}(\pi)\boldsymbol{X}_i\|^2 + \sum_{j=0}^{d+1} |\mathcal{I}_\ell|\lambda\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\|_1\right) \\ &> \max_{\pi \in \Pi_0}\left(\sum_{i \in \mathcal{I}_\ell} \|\boldsymbol{X}_i - \boldsymbol{W}(\pi)\boldsymbol{X}_i\|^2 + \sum_{j=0}^{d+1} |\mathcal{I}_\ell|\lambda\|\boldsymbol{W}_j(\pi)\|_1\right).\end{aligned} \tag{S43}$$

By definition, this implies that $\widetilde{\boldsymbol{W}}^{(\ell)} = \widetilde{\boldsymbol{W}}^{(\ell)}(\pi^*)$ for some true ordering $\pi^* \in \Pi^*$. This proves the first assertion of Proposition 1. In the third step, we derive the convergence rate of $\widetilde{\boldsymbol{W}}^{(\ell)}$, and completes the second assertion of Proposition 1.

Step 1: For any π, i, j, let $\varepsilon_{i,j}(\pi) = X_{i,j} - \boldsymbol{X}_i^\top \boldsymbol{W}_j(\pi)$ denote the residual. By definition,

$$\begin{aligned}&\sum_{j=0}^{d+1}\sum_{i \in \mathcal{I}_\ell} |X_{i,j} - \boldsymbol{X}_i^\top \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)|^2 + \lambda|\mathcal{I}_\ell|\sum_{j=0}^{d+1} \|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\|_1 \qquad \text{(S44)} \\ \leq &\sum_{j=0}^{d+1}\sum_{i \in \mathcal{I}_\ell} |X_{i,j} - \boldsymbol{X}_i^\top \boldsymbol{W}_j(\pi)|^2 + \lambda|\mathcal{I}_\ell|\sum_{j=0}^{d+1} \|\boldsymbol{W}_j(\pi)\|_1 = \sum_{j=0}^{d+1}\sum_{i \in \mathcal{I}_\ell} |\varepsilon_{i,j}(\pi)|^2 + \lambda|\mathcal{I}_\ell|\sum_{j=0}^{d+1} \|\boldsymbol{W}_j(\pi)\|_1.\end{aligned}$$

Note that

$$\begin{aligned}&\sum_{i \in \mathcal{I}_\ell} |X_{i,j} - \boldsymbol{X}_i^\top \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)|^2 = \sum_{i \in \mathcal{I}_\ell} |\varepsilon_{i,j}(\pi) + \boldsymbol{X}_i^\top \boldsymbol{W}_j(\pi) - \boldsymbol{X}_i^\top \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)|^2 \\ = &\sum_{i \in \mathcal{I}_\ell} |\varepsilon_{i,j}(\pi)|^2 + 2\sum_{i \in \mathcal{I}_\ell} \varepsilon_{i,j}(\pi)\left\{\boldsymbol{X}_i^\top \boldsymbol{W}_j(\pi) - \boldsymbol{X}_i^\top \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\} + \sum_{i \in \mathcal{I}_\ell}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi) - \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2.\end{aligned}$$

Let $s_j(\pi)$ and $\tilde{s}_j^{(\ell)}(\pi)$ denote the number of nonzero elements in $\boldsymbol{W}_j(\pi)$ and $\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)$, respectively. By Theorem 7.1 of van de Geer and Bühlmann (2013), we have,

$$
\begin{aligned}
\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell} 2\left|\varepsilon_{i,j}(\pi)\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right| - \delta\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}\left|\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right|^2 & \\
\leq \sum_{j=0}^{d+1}\frac{\kappa\left\{s_j(\pi)+\tilde{s}_j^{(\ell)}(\pi)\right\}\log n}{\delta}&,
\end{aligned} \tag{S45}
$$

for some constant $\kappa > 0$, any $\delta > 0$, and any order π. Under the event defined in (S45),

$$
\begin{aligned}
\sum_{i\in\mathcal{I}_\ell}\left|X_{i,j}-\boldsymbol{X}_i^\top\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right|^2 \geq & \sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2 \\
& + (1-\delta)\sum_{i\in\mathcal{I}_\ell}\sum_{j=0}^{d+1}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2 - \sum_{j=0}^{d+1}\frac{\kappa\left\{s_j(\pi)+\tilde{s}_j^{(\ell)}(\pi)\right\}\log n}{\delta}.
\end{aligned} \tag{S46}
$$

Setting $\delta = 1/2$, together with (S44), we have,

$$
\begin{aligned}
\frac{1}{2}\sum_{i\in\mathcal{I}_\ell}\sum_{j=0}^{d+1}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2 + \lambda|\mathcal{I}_\ell|\sum_{j=0}^{d+1}\left\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_1 & \\
\leq \sum_{j=0}^{d+1}2\kappa\left\{s_j(\pi)+\tilde{s}_j^{(\ell)}(\pi)\right\}\log n + \lambda|\mathcal{I}_\ell|\sum_{j=0}^{d+1}\left\|\boldsymbol{W}_j(\pi)\right\|_1&.
\end{aligned}
$$

Using similar arguments in bounding $\max_{j_1,j_2,\ell} I_{N,1}(j_1,j_2,\ell)$ in Step 1 of the proof of Theorem 1, we can show there exists constants $\kappa^* \leq \lambda_{\min}(\boldsymbol{\Sigma}_0) \leq \lambda_{\max}(\boldsymbol{\Sigma}_0) \leq 1/\kappa^*$ such that

$$
\begin{aligned}
\kappa^*|\mathcal{I}_\ell|\sum_{j=0}^{d+1}\left\|\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_2^2 \leq & \frac{1}{2}\sum_{i\in\mathcal{I}_\ell}\sum_{j=0}^{d+1}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2 \\
\leq & \frac{1}{\kappa^*}|\mathcal{I}_\ell|\sum_{j=0}^{d+1}\left\|\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_2^2,
\end{aligned} \tag{S47}
$$

for any order π. It then follows that, for any order π,

$$
\begin{aligned}
\kappa^*\sum_{j=0}^{d+1}\left\|\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_2^2 + \lambda\sum_{j=0}^{d+1}\left\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_1 \leq & \sum_{j=0}^{d+1}\frac{4c}{n}\left\{s_j(\pi)+\tilde{s}_j^{(\ell)}(\pi)\right\}\log n \\
& + \lambda\sum_{j=0}^{d+1}\|\boldsymbol{W}_j(\pi)\|_1.
\end{aligned} \tag{S48}
$$

Next, define the $(d+2)$-dimensional vector,

$$\boldsymbol{\mu}(\pi) = \left\{ \widetilde{\boldsymbol{W}}_0^{(\ell)\top}(\pi) - \boldsymbol{W}_0^{(\ell)\top}(\pi), \widetilde{\boldsymbol{W}}_1^{(\ell)\top}(\pi) - \boldsymbol{W}_1^{(\ell)\top}(\pi), \ldots, \widetilde{\boldsymbol{W}}_{d+1}^{(\ell)\top}(\pi) - \boldsymbol{W}_{d+1}^{(\ell)\top}(\pi) \right\}^\top .$$

Let $\mathcal{M}(\pi)$ denote the support of $\boldsymbol{\mu}(\pi)$. Our goal is to bound $\|\boldsymbol{\mu}(\pi)\|_2 = \sum_{j=0}^{d+1} \|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi) - \boldsymbol{W}_j(\pi)\|_2^2$. We next consider two cases.

First, when $\|\boldsymbol{\mu}(\pi)\|_2 \leq \sum_{j=0}^{d+1} 4c \left\{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \right\} \log n/(n\kappa^*)$, it is already bounded.

Second, when $\|\boldsymbol{\mu}(\pi)\|_2 > \sum_{j=0}^{d+1} 4c \left\{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \right\} \log n/(n\kappa^*)$, we aim to obtain a larger upper bound. Specifically, under the event defined in (S48), we have, for all π,

$$\sum_{j=0}^{d+1} \left\| \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi) \right\|_1 \leq \sum_{j=0}^{d+1} \left\| \boldsymbol{W}_j(\pi) \right\|_1 .$$

It follows that, for all π,

$$\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_1 \leq \|\boldsymbol{\mu}(\pi)_{\mathcal{M}^c(\pi)}\|_1, \tag{S49}$$

where $\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}$ and $\boldsymbol{\mu}(\pi)_{\mathcal{M}^c(\pi)}$ denote the sub-vector formed by the elements in $\mathcal{M}(\pi)$ and $\mathcal{M}^c(\pi)$, respectively. Then, under the event defined in (S49), and by the definition of s_0,

$$\begin{aligned} \sum_{j=0}^{d+1} \left\| \boldsymbol{W}_j(\pi) \right\|_1 - \sum_{j=0}^{d+1} \left\| \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi) \right\|_1 &\leq \|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_1 + \|\boldsymbol{\mu}(\pi)_{\mathcal{M}^c(\pi)}\|_1 \leq 2\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_1 \\ &\leq 2\sqrt{|\mathcal{M}(\pi)|}\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2 \leq 2\sqrt{s_0(d+2)}\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2. \end{aligned}$$

By (S48), we have that, for all π,

$$\begin{aligned} \kappa^* \|\boldsymbol{\mu}(\pi)\|_2^2 &\leq \sum_{j=0}^{d+1} \frac{4c}{n} \left\{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \right\} \log n + 2\lambda\sqrt{s_0(d+2)}\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2, \\ \kappa^* \|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2^2 &\leq \sum_{j=0}^{d+1} \frac{4c}{n} \left\{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \right\} \log n + 2\lambda\sqrt{s_0(d+2)}\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2. \end{aligned}$$

It then follows that, for all π and some positive constant $O(1)$,

$$\|\boldsymbol{\mu}(\pi)_{\mathcal{M}(\pi)}\|_2 \leq O(1) \left(\lambda\sqrt{s_0(d+2)} + \sqrt{\sum_{j=0}^{d+1} \frac{4c}{n} \left\{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \right\} \log n} \right).$$

Under the event defined in (S49), and by Equations (A5), (A6) in Zhou (2009), we have,

$$\|\boldsymbol{\mu}(\pi)\|_2 \leq O(1) \left(\lambda\sqrt{s_0(d+2)} + \sqrt{\sum_{j=0}^{d+1} \frac{4c}{n} \{ s_j(\pi) + \tilde{s}_j^{(\ell)}(\pi) \} \log n} \right), \tag{S50}$$

for all π.

Therefore, in both cases, we show that $\|\boldsymbol{\mu}(\pi)\|_2$ is uniformly bounded for all π. This completes Step 1.

Step 2: Under the events defined in (S47) and (S50), we have that, for all π,

$$\sum_{i\in\mathcal{I}_\ell}\sum_{j=0}^{d+1}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi)-\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2 \le O(1)d^2\log n.$$

Combining this together with (S46) yields that, for all π,

$$\sum_{i\in\mathcal{I}_\ell}\left|X_{i,j}-\boldsymbol{X}_i^\top\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right|^2 \ge \sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2+O(1)d^2\log n. \tag{S51}$$

Under the omega-min condition in (A6), using similar arguments in the proof of Lemma 7.8 of van de Geer and Bühlmann (2013), we have,

$$\min_{\pi\notin\Pi_0}\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2 \ge \max_{\pi\in\Pi_0}\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2+\kappa(d+2)n,$$

for some constant $\kappa>0$. Under the event defined in (S51), we have,

$$\begin{aligned}\min_{\pi\notin\Pi_0}\left\{\sum_{i\in\mathcal{I}_\ell}\|\boldsymbol{X}_i-\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)\boldsymbol{X}_i\|^2+\sum_{j=0}^{d+1}|\mathcal{I}_\ell|\lambda\left\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\|_1\right\} &\ge \min_{\pi\notin\Pi_0}\sum_{i\in\mathcal{I}_\ell}\left\|\boldsymbol{X}_i-\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)\boldsymbol{X}_i\right\|^2\\ &\ge \max_{\pi\in\Pi_0}\sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2+\kappa(d+2)n-O(1)d^2\log n.\end{aligned}$$

Under the Condition $\|\boldsymbol{W}_0\|_2=O(1)$ for any $\pi\in\Pi^*$, we have $\|\boldsymbol{W}_j(\pi)\|_2=\|\boldsymbol{W}_{0,j}\|_2=O(1)$. As a result,

$$\max_{\pi\in\Pi_0}\sum_{j=0}^{d+1}\lambda|\mathcal{I}_\ell|\|\boldsymbol{W}_j(\pi)\|_1=\sum_{j=0}^{d+1}\lambda|\mathcal{I}_\ell|\|\boldsymbol{W}_{0,j}\|_1\le\sqrt{d+2}\sum_{j=0}^{d+1}|\mathcal{I}_\ell|\lambda\|\boldsymbol{W}_{0,j}\|_2=O\left(d^{3/2}n^{1/2}\sqrt{\log n}\right).$$

It follows that,

$$\begin{aligned}&\min_{\pi\notin\Pi_0}\left(\sum_{i\in\mathcal{I}_\ell}\|\boldsymbol{X}_i-\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)\boldsymbol{X}_i\|^2+\sum_{j=0}^{d+1}|\mathcal{I}_\ell|\lambda\|\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\|_1\right)\\ \ge\ &\max_{\pi\in\Pi_0}\sum_{j=0}^{d+1}\left(\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2+\sum_{j=0}^{d+1}|\mathcal{I}_\ell|\lambda\|\boldsymbol{W}_j(\pi)\|_1\right)\\ &+\kappa(d+2)n-O(1)d^2\log n-O\left(d^{3/2}n^{1/2}\sqrt{\log n}\right),\end{aligned}$$

where the last line is strictly positive. As a result, we obtain (S43). This completes Step 2.

Step 3: Built on the results obtained in Step 2, we have $\widehat{\pi} \in \Pi_0$, where $\widehat{\pi}$ is the estimated ordering. To complete the proof, it suffices to show that $\max_{\pi\in\Pi_0}\max_{j\in\{0,\ldots,d+1\}}\|\widetilde{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_2 = O\left(n^{-1/2}\sqrt{s_0\log n}\right)$. Since the number of elements in Π^* is bounded by $O(n^{\kappa^*})$ for some $\kappa^* > 0$, by Bernstein's inequality and Bonferroni's inequality, we have that,

$$\max_{\pi\in\Pi_0}\max_{k\in\{0,1,\ldots,j-1\}}|\sum_{i\in\mathcal{I}_\ell}\varepsilon_{i,j}(\pi)X_{i,\pi_k}| = O\left(n^{1/2}\sqrt{\log n}\right).$$

The explicit convergence rate of $\max_{\pi\in\Pi_0}\max_{j\in\{0,\ldots,d+1\}}\|\widetilde{\boldsymbol{W}}_j^{(\ell)} - \boldsymbol{W}_{0,j}\|_2$ can then be similarly derived following Theorem 3.1 of Zhou (2009). This completes Step 3.

Remark: Finally, we discuss the possibility of relaxing the condition $d = o(n)$, by imposing some sparsity conditions on $\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)$, and the population limit of $\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)$, i.e., $\boldsymbol{W}_0(\pi)$. In this case, the condition $d \ll n$ can be relaxed to $d = O(n^{\kappa_1})$, for some constant $\kappa_1 > 0$. Since we do not require $\kappa_1 < 1$, the dimension d can grow at a much faster rate than n.

Specifically, suppose the number of nonzero elements in each column of $\widetilde{\boldsymbol{W}}^{(\ell)}(\pi)$ and $\boldsymbol{W}_0(\pi)$ satisfies that $\max_{j,\pi}[\max\{s_j(\pi), \tilde{s}_j^{(\ell)}(\pi)\}] = O(n^{\kappa_*})$, for some constant $0 < \kappa_* < 1$. It follows from (S46) and (S50) that

$$\begin{aligned}\sum_{i\in\mathcal{I}_\ell}\left|X_{i,j} - \boldsymbol{X}_i^\top\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right|^2 \geq \sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2 + (1-\delta)\sum_{i\in\mathcal{I}_\ell}\sum_{j=0}^{d+1}\left[\boldsymbol{X}_i^\top\left\{\boldsymbol{W}_j(\pi) - \widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right\}\right]^2 \\ - \sum_{j=0}^{d+1}\frac{O(1)n^{\kappa_*}\log n}{\delta}, \\ \|\boldsymbol{\mu}(\pi)\|_2 \leq O(1)\left(\lambda\sqrt{d}n^{\kappa_*/2} + \sqrt{d}n^{\kappa_*/2-1/2}\sqrt{\log n}\right),\end{aligned}$$

for all π, where $O(1)$ denotes some positive constant. Therefore, we have

$$\sum_{i\in\mathcal{I}_\ell}\left|X_{i,j} - \boldsymbol{X}_i^\top\widetilde{\boldsymbol{W}}_j^{(\ell)}(\pi)\right|^2 \geq \sum_{j=0}^{d+1}\sum_{i\in\mathcal{I}_\ell}|\varepsilon_{i,j}(\pi)|^2 + O(1)dn^{\kappa_*}\log n,$$

for all π. The rest of the proof follows using similar arguments as in Steps 2 and 3. □

S10 Additional numerical results

First, we provide more details about the simulation setup. Table S1 reports the corresponding mediators with nonzero mediation effects, and their associated $\delta(q)$, where

Scenario A: $(d, p_1, p_2) = (50, 0.05, 0.15)$, $n = 100, 200$

q	10	12	20	28	30	41
$\delta(q)$	1.06	1.03	0.63	1.08	0.64	1.31

Scenario B: $(d, p_1, p_2) = (100, 0.025, 0.075)$, $n = 250, 500$

q	5	14	17	20	30	71	80	82	89
$\delta(q)$	0.99	0.33	0.92	0.85	0.85	1.37	0.72	0.72	0.72

Scenario C: $(d, p_1, p_2) = (150, 0.02, 0.05)$, $n = 250, 500$

q	19	23	51	52	54	56	60	81	92	93
$\delta(q)$	1.24	0.93	1.71	1.28	1.90	1.90	0.94	0.88	0.88	1.02

Table S1: Mediators with nonzero mediation effects, $q = 1, \dots, d$.

$\delta(q) = (\boldsymbol{W}_0^*)_{d+1,q}(\boldsymbol{W}_0^*)_{q,0}$, $\boldsymbol{W}_0^*$ is constructed based on $\boldsymbol{W}_0$, $q = 1, \dots, d$. Meanwhile, Figure S3 shows an example of the adjacent matrix when $d = 100$.

Next, we report the simulation results for testing of a single mediator. Figures 2 and S5 show the results for $d = 100$ and $d = 150$, respectively, which complement the results for $d = 50$ reported in Section 6 of the paper. We observe a similar qualitative pattern that our proposed test achieves a valid size under the null hypothesis, and achieves a larger empirical power under the alternative hypothesis.

Next, we report the simulation results for multiple testing. Figure S6 show the results for $d = 100$ and $d = 150$, respectively, which complement the results for $d = 50$ reported in Section 6 of the paper. We again observe a similar qualitative pattern that both our test and the standard Benjamini-Yekutieli (BY) procedure achieve a valid false discovery control, whereas our method is more powerful than BY.

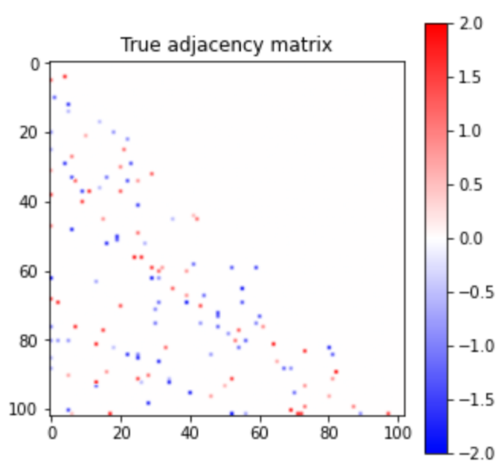

Figure S3: The example adjacent matrix with $d = 100$.

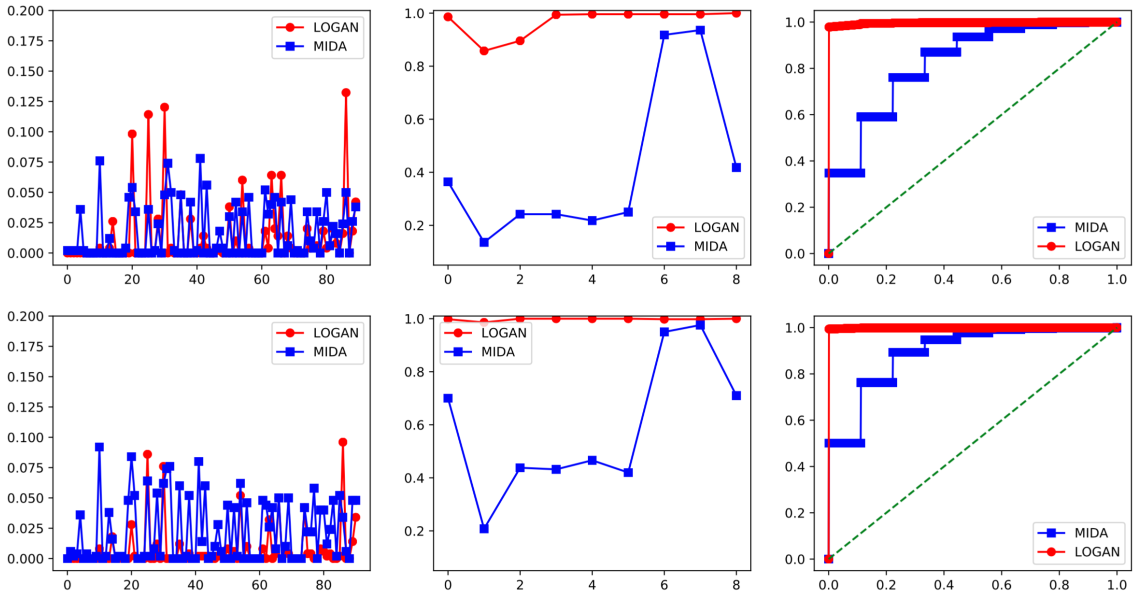

Figure S4: Empirical rejection rate and ROC curve of the proposed test, LOGAN, and the test of Chakrabortty et al. (2018), MIDA, when $d = 100$. The upper panels: $n = 250$, and the bottom panels: $n = 500$. The left panels: under H_0, the middles panels: under H_1, where the horizontal axis is the mediator index, and the right panels: the average ROC curve.

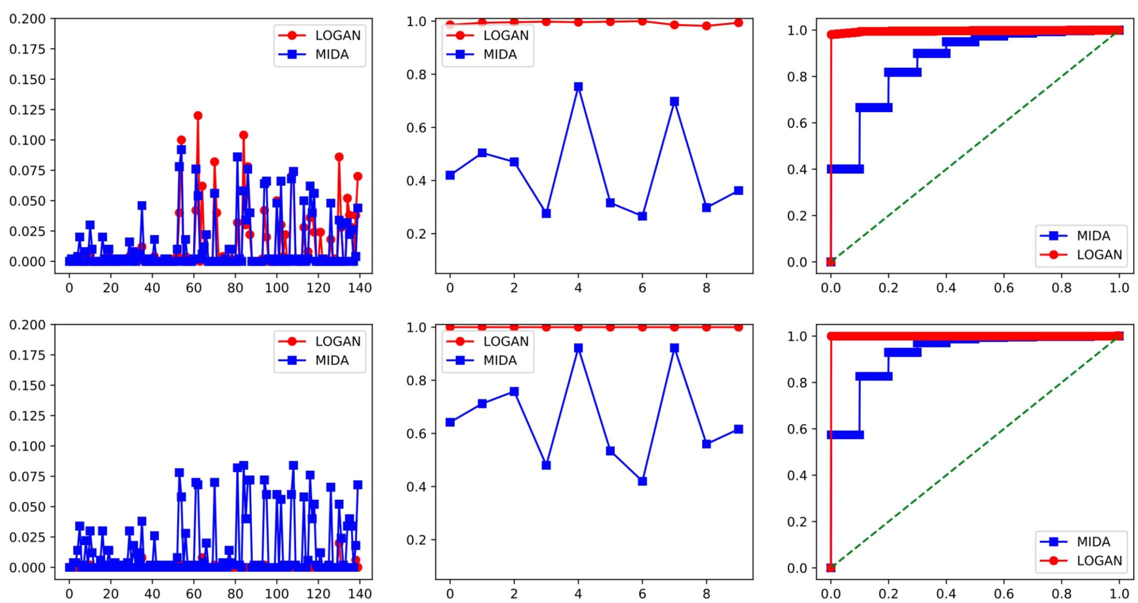

Figure S5: Empirical rejection rate and ROC curve of the proposed test, LOGAN, and the test of Chakrabortty et al. (2018), MIDA, when $d = 150$. The upper panels: $n = 250$, and the bottom panels: $n = 500$. The left panels: under H_0, the middles panels: under H_1, where the horizontal axis is the mediator index, and the right panels: the average ROC curve.

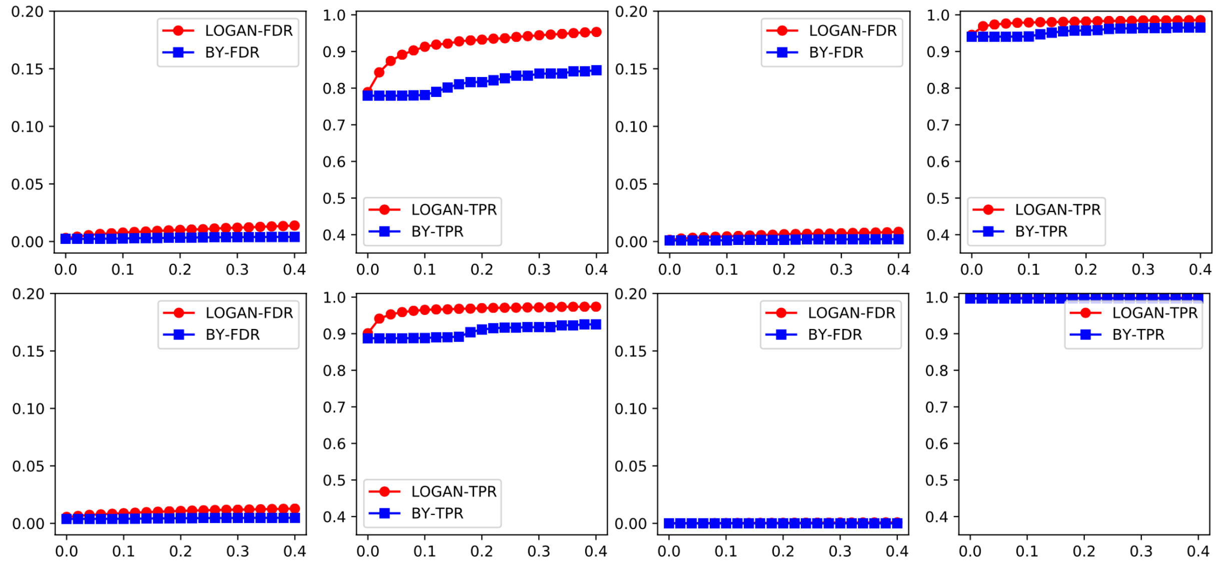

Figure S6: False discover rate and true positive rate of the proposed method and the Benjamini-Yekutieli procedure. The horizontal axis corresponds to the significance level α. The upper panels: $d = 100$, and the bottom panels: $d = 150$. The left two panels: $n = 250$, and the right two panels: $n = 500$.

Finally, we evaluate the constant variance condition for the real data example. Following a similar idea of Li et al. (2019), we compute and plot the residuals $X_{i,j} - \widetilde{\boldsymbol{W}}^{(\ell)} \boldsymbol{X}_i$, for $i \in \mathcal{I}_\ell^c$, $j \in \{0, \ldots, d+1\}$, $\ell = 1, 2$. Figure S7 shows the boxplots of such residuals for the amyloid negative and amyloid positive groups, respectively. It is seen that the lengths of boxes and whiskers are similar across different variables, suggesting no strong evidence against the constant variance condition.

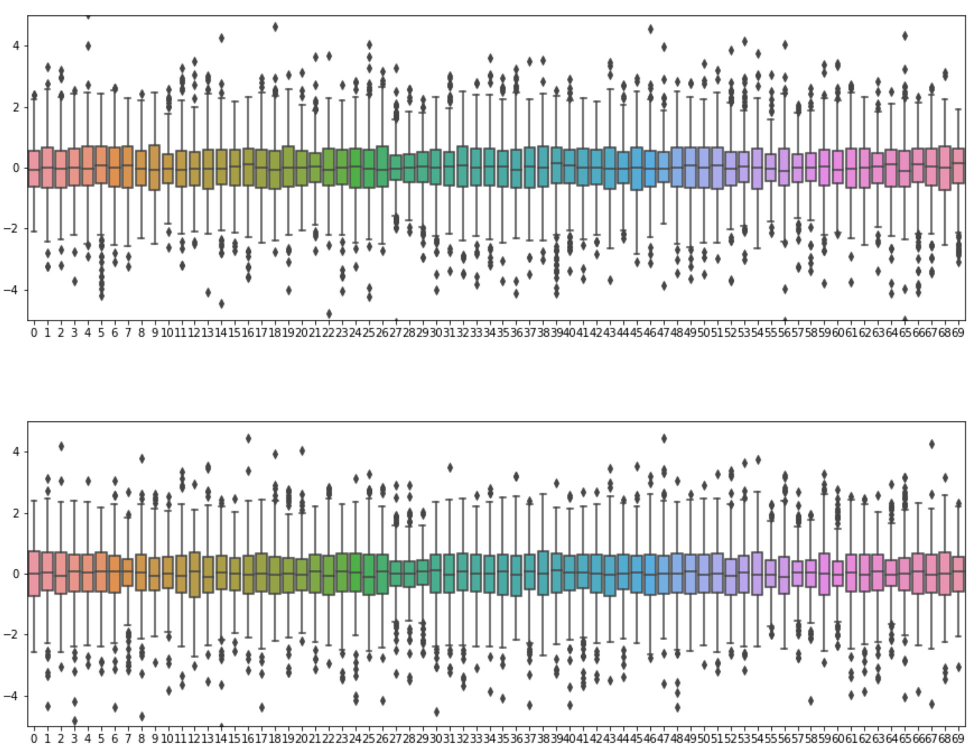

Figure S7: Residual plots for model diagnosis. The top panel: the amyloid negative group, and the bottom panel: the amyloid positive group.

References

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Chakrabortty, A., Nandy, P., and Li, H. (2018). Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162(1-2):47–70.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Detailed proof of nazarov's inequality. *arXiv preprint arXiv:1711.10696*.

Djordjilović, V., Page, C. M., Gran, J. M., Nø st, T. H., Sandanger, T. M., Veierø d, M. B., and Thoresen, M. (2019). Global test for high-dimensional mediation: testing groups of potential mediators. *Statistics in Medicine*, 38(18):3346–3360.

Li, C., Shen, X., and Pan, W. (2019). Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, 0(0):1–16.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

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.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Shi, C., Fan, A., Song, R., and Lu, W. (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *The Annals of Statistics*, 46(3):925–957.

Shi, C., Rui, S., Lu, W., and Li, R. (2020). Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association*, accepted.

van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567.

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.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv: 0912.4045*.