Web-based Supplementary Materials for Covariate selection with group lasso and doubly robust estimation of causal effects by Koch, Vock, and Wolfson

April 23, 2017

Web Appendix A

Here, we present conditions under which the simultaneous variable selection problem defined by Equation (3) in Section 3.2 in the main paper has a unique solution. An immediate corollary is that a solution exists when Φ_{sum} is given by a sum of the squared error and logistic loss, i.e., when defining linear and logistic regression models for the outcome and treatment, respectively.

For notational purposes, let $\mathbf{D}$ denote the working data $\{Y, A, \mathbf{V}\}$ and $L(\beta|\mathbf{D})$ be the empirical loss, i.e.,

$$L(\beta|\mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n} \Phi_{sum}(Y_i, A_i, \mathbf{V}_i; \beta). \tag{1}$$

Yang and Zou (2015) show that Equation (3) in the main paper has a solution provided the loss function Φ_{sum} satisfies a so-called quadratic majorization (QM) condition, i.e., if and only if the following two assumptions hold:

(i) $L(\beta|\mathbf{D})$ is a differentiable function of β, i.e., $\nabla L(\beta, \mathbf{D})$ exists everywhere.

(ii) There exists a p x p matrix $\mathbf{H}$, which may only depend on $\mathbf{D}$, such that for all β, β^*

$$L(\beta|\mathbf{D}) \leq L(\beta^*|\mathbf{D}) + (\beta - \beta^*)^T \nabla L(\beta^*|\mathbf{D}) + \frac{1}{2}(\beta - \beta^*)^T \mathbf{H}(\beta - \beta^*)$$

We state and prove the following extension to their result which characterizes a class of loss functions of the form of Equation (2) in the main paper that satisfy the QM condition:

Lemma 1 *Let $\Phi_{sum}(Y, A, f, g) = \Phi_{out}(Y, f) + \Phi_{trt}(A, g)$, where Φ_{out} is the loss function used to link outcome Y with predictors $Z_1 = \{Z_{11}, \ldots, Z_{1r}\}$ through a linear predictor $f = \alpha^T Z_1$, and Φ_{trt} is the loss function used to link treatment A with predictors $Z_2 = \{Z_{21}, \ldots, Z_{2s}\}$ through a linear predictor $g = \gamma^T Z_2$. Let $Z = \{Z_{11}, \ldots, Z_{1r}, Z_{21}, \ldots, Z_{2s}\}$. Assume Φ_{out} is differentiable with respect to the coefficient parameters in f and write $\Phi'_{out} = \frac{\partial \Phi_{out}(Y,f)}{\partial f}$, and similarly, assume Φ_{trt} is differentiable with respect to the coefficient parameters in g and write $\Phi'_{trt} = \frac{\partial \Phi_{trt}(A,g)}{\partial g}$. Then:*

(1). If Φ'_{out} and Φ'_{trt} are Lipschitz continuous with constants C_1 and C_2 such that

$$(i)\ |\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| \leq C_1 |f_1 - f_2|\ \forall Y, f_1, f_2,$$

and

$$(ii)\ |\Phi'_{trt}(A, g_1) - \Phi'_{trt}(A, g_2)| \leq C_2 |g_1 - g_2|\ \forall A, g_1, g_2,$$

then the QM condition holds for Φ_{sum} and $\boldsymbol{H} = \frac{2(C_1+C_2)}{n}\mathbf{Z}^T\mathbf{Z}$.

(2). If $\Phi''_1 = \frac{\partial \Phi^2(Y,f)}{\partial f^2}$ and $\Phi''_2 = \frac{\partial \Phi^2(A,g)}{\partial g^2}$ exist and there are constants C_3 and C_4 such that

$$(i)\ \Phi''_1 \leq C_3\ \forall Y, f,$$

and

$$(ii)\ \Phi''_2 \leq C_4\ \forall A, g$$

then the QM condition holds for Φ_{sum} *and* $\boldsymbol{H} = \frac{C_3+C_4}{n}\mathbf{Z}^T\mathbf{Z}$.

(3) If

(i) Φ_{out} *satisfies condition (1)(i) with constant* C_1,

(ii) Φ_{trt} *satisfies condition (2)(ii) with constant* C_2,

and

(iii) $\Phi_2'' = \frac{\partial\Phi^2(A,g)}{\partial g^2} \geq C_L \; \forall A, g$ *(i.e.,* Φ_2'' *is bounded),*

or

(i) Φ_{trt} *satisfies condition (1)(ii) with constant* C_1,

(ii) Φ_{out} *satisfies condition (2)(i) with constant* C_2,

and

(iii) $\Phi_1'' = \frac{\partial\Phi^2(Y,f)}{\partial f^2} \geq C_L \; \forall Y, f$ *(i.e.,* Φ_1'' *is bounded),*

then the QM condition holds for Φ_{sum} *and* $\boldsymbol{H} = \frac{2(C_1+C_2^*)}{n}\mathbf{Z}^T\mathbf{Z}$, *where* $C_2^* = \max\{|C_2|, |C_L|\}$.

Proof. Before proving Lemma 1, we first present a lemma (without observation weights) from Yang and Zou (2015):

Lemma 2 *Assume* $\Phi(y, f)$ *is differentiable with respect to* f *and write* $\Phi'_f = \frac{\partial\Phi(y,f)}{\partial f}$. *Then*

$$\nabla\Phi(\beta|\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n}\tau_i\Phi'(y_i, x_i^T\beta)x_i$$

(1). If Φ'_f *is Lipschitz continuous with constant* C *such that*

$$|\Phi'_f(y, f_1) - \Phi'_f(y, f_2)| \leq C|f_1 - f_2| \; \forall y, f_1, f_2,$$

then the QM condition holds for Φ *and* $\boldsymbol{H} = \frac{2C}{n}\mathbf{X}^T\mathbf{X}$.

(2). If $\Phi''_f = \frac{\partial\Phi^2(y,f)}{\partial f^2}$ *exists and* $\Phi''_f \leq C_2 \; \forall y, f$,

then the QM condition holds for Φ *and* $\boldsymbol{H} = \frac{C_2}{n}\mathbf{X}^T\mathbf{X}$.

Proving (1): We have

$$|\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| \leq C_1|f_1 - f_2| \; \forall Y, f_1, f_2,$$

and

$$|\Phi'_{trt}(A, g_1) - \Phi'_{trt}(A, g_2)| \leq C_2|g_1 - g_2| \; \forall A, g_1, g_2$$

Then condition (1) in Lemma 2 is satisfied for the outcome and treatment loss functions with constants C_1 and C_2, respectively. This implies

$$|\Phi'_{sum}(Y, A, f_1, f_1) - \Phi'_{sum}(Y, A, f_2, f_2)|$$

$$= |\Phi'_{out}(Y, f_1) + \Phi'_{trt}(A, f_1) - \Phi'_{out}(Y, f_2) - \Phi'_{trt}(A, f_2)|$$

$$\leq |\Phi'_{out}(Y, f_1) - \Phi'_{out}(Y, f_2)| + |\Phi'_{trt}(A, f_1) - \Phi'_{trt}(A, f_2)|$$

$$\leq C_1|f_1 - f_2| + C_2|f_1 - f_2| \; \forall Y, A, f_1, f_2.$$

To prove (2) in Lemma 1: We have constants C_3 and C_4 such that $\Phi''_{out}(Y, f) \leq C_3$ and $\Phi''_{trt}(A, g) \leq C_4$ for all Y, f, g. Then

$$\Phi''_{sum}(Y, A, f, g) = \Phi''_{out}(Y, f) + \Phi''_{trt}(A, g) \leq C_3 + C_4.$$

Finally, to prove (3) in Lemma 1: Assume condition (1) in Lemma 2 is satisfied for, say (WLOG), Φ_{out} with constant C_1, and also assume Φ_{trt} satisfies condition (2) in Lemma 2 with constant C_2 such that $\Phi''_{trt} \geq C_L$. Then since Φ''_{trt} is bounded, we know Φ'_{trt} is Lipschitz continuous with constant C_2^* (bounded derivative implies Lipschitz continuity). The proof then concludes following the proof of (1) with constants C_1 and C_2^*. □

To use linear and logistic regression to model the outcome and treatment, respectively,

and (naturally) letting Φ_{out} be the squared-error loss function and Φ_{trt} be the loss function proportional to the binomial log-likelihood, we have $\Phi''_{ls} = 1$ and $\Phi''_{logit} \leq \frac{1}{4}$, meaning the QM condition holds for $\mathbf{H} = \frac{(5/4)}{n}\mathbf{Z}^T\mathbf{Z}$ by Lemma 1 condition (2).

When the QM condition is met (i.e., when the conditions of Lemma 1 are satisfied), we are able to solve for β in Equation (3) in the main paper using the groupwise-majorization-descent (GMD) algorithm (for details, see Yang and Zou (2015)), a computationally efficient and unified algorithm allowing for general design matrices.

Web Appendix B

Here, we provide a proof of Theorem 1 in the main text, which is re-stated here:

Theorem 1 *Assume the number of covariates p and sample size n are such that $\frac{\log(2p)}{n} \leq 1$. Also assume the Group Stabil condition is satisfied with $c_0 = 3$ and $\epsilon = \frac{1}{2n}$. Let $\zeta^* = 2\sum_{g=1}^{p} I(\alpha_g^* \neq 0)$. Then, for sufficiently large λ_n and with high probability, we have*

$$\sum_{g=1}^{p} \left\| \left(\hat{\beta}_g - \beta_g^*\right) I(\alpha_g^* \neq 0) \right\|_2 \leq \frac{\underset{g\in\{1,\ldots,p\}}{max}\{|v_g|\}}{\sqrt{2}} \left(\frac{4}{c_n k}\lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right)\frac{1}{2n} \right)$$

where $0 < k < 1$ is defined in Definition 1 (pg. 12), and

$$c_n = \min_{[|x|<L(9B+\frac{1}{n})]\cap\Theta} \left\{ \frac{\Psi''_{out}(x) + \Psi''_{trt}(x)}{2} \right\}.$$

We recall that λ_n is the penalty, $\hat{\beta}$ is the group lasso estimator in our set-up, β^* is the vector of true/least false coefficient parameters in the outcome and treatment models, and β_g^* is the sub-vector of β^* associated with group g (in our case, covariate g). We let p^* be the total number of columns in the design matrices of the outcome and treatment models (i.e, p^* is the length of β^*), which we denote by $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. We assume $(Z_{out,i}, Z_{trt,i}, Y_i, A_i)$ are i.i.d. copies of (Z_{out}, Z_{trt}, Y, A) for $i = 1, \ldots, n$, where $Y|Z_{out}$ and $A|Z_{trt}$ are modeled

by distributions F_{out} and F_{trt} both on $\mathbb{R}$ and from the exponential family, respectively, and $Z_{out,i}$ and $Z_{trt,i}$ are the ith rows of $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. The natural parameter space is denoted by $\Theta := \Theta_{out} \cup \Theta_{trt}$, where

$$\Theta_{out} = \left\{ \theta \in \mathbb{R} : \int \exp(\theta x) F_{out}(dx) < \infty \right\}$$

and

$$\Theta_{trt} = \left\{ \theta \in \mathbb{R} : \int \exp(\theta x) F_{trt}(dx) < \infty \right\}.$$

L and B apply to assumptions (H.1–3):

(H.1): the pair of variables (Z_{out}, Z_{trt}) are almost surely bounded by a constant L, i.e., there exists a constant $L > 0$ such that
$||(Z_{out}, Z_{trt})||_\infty \leq L$ a.s.

(H.2): for all $x \in [-L, L]^{p^*}$, $\beta^{*T} x \in \text{Int}(\Theta)$

(H.3): There exists a constant $B > 0$ such that $\sum_{g=1}^{G_n} \sqrt{d_g} ||\beta_g^*||_2 \leq B$

We consider $\Lambda = \left\{ \beta \in \mathbb{R}^{p^*} : \forall x \in [-L, L]^{p^*}, \beta^T x \in \Theta \right\}$. We define d_g to be the size of group g, $g \in \{1, \ldots, G_n\}$, and let $d_{min} := \min_{g \in \{1,\ldots,G_n\}} d_g$ and $d_{max} := \max_{g \in \{1,\ldots,G_n\}} d_g$ denote the smallest and largest group sizes, respectively. Letting $\Phi_{sum}(\beta) = \Phi_{sum}(Y, A, \mathbf{Z}; \beta)$, the empirical process $(\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta))$ can be written as:

$$\begin{aligned}
(\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta)) &= (\mathbb{P}_n - \mathbb{P})[\Phi_{out}(\beta) + \Phi_{trt}(\beta)] \\
&= (\mathbb{P}_n - \mathbb{P})[\Phi_{out,l}(\beta)] + (\mathbb{P}_n - \mathbb{P})[\Phi_{out,\Psi}(\beta)] \\
&+ (\mathbb{P}_n - \mathbb{P})[\Phi_{trt,l}(\beta)] + (\mathbb{P}_n - \mathbb{P})[\Phi_{trt,\Psi}(\beta)],
\end{aligned}$$

where $\Phi_{out,l} = -Y\alpha'\mathbf{Z}_{out}$, $\Phi_{out,\Psi} = \Psi_{out}(\alpha'\mathbf{Z}_{out})$, $\Phi_{trt,l} = -A\gamma'\mathbf{Z}_{trt}$, and $\Phi_{trt,\Psi} = \Psi_{trt}(\gamma'\mathbf{Z}_{trt})$; $\Psi''_{out}(x)$ and $\Psi''_{trt}(x)$ in Theorem 1 denote the second derivatives of $\Psi_{out}(x)$ and $\Psi_{trt}(x)$, respectively. $\Phi_{.,l}$ is used to denote the *linear* part of Φ and $\Phi_{.,\Psi}$ is used to denote the

part which depends on the link function between the canonical parameter and the linear predictor. For example, if modeling the outcome with linear regression (i.e., using squared error loss), then $\Phi_{out,\Psi} = \alpha' \mathbf{Z}_{out}$, and if modeling the treatment with logistic regression, $\Phi_{trt,\Psi} \propto n \log(1 + \exp(\gamma' \mathbf{Z}_{trt})))$.

We define

$$L_g := \left\| \frac{\hat{\beta}_g^{ls}}{\sqrt{d_g}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g \right)^T - E \left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g \right)^T \right\} \right\|_2$$

for all $g \in \{1, \ldots, G_n\}$, where $Z_{out,i}^g$ and $Z_{trt,i}^g$ denote the elements on the ith rows and gth columns of $\mathbf{Z}_{out}$ and $\mathbf{Z}_{trt}$, respectively. We also define

$$\mathcal{A} = \bigcap_{g=1}^{G_n} \left\{ L_g \leq \frac{\lambda_n}{2} \right\}$$

and

$$\mathcal{B} = \left\{ \sup_{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g - \beta_g^*||_2 \leq M} |v_n(\beta, \beta^*)| \leq \frac{\lambda_n}{2} \right\}$$

where

$$v_n = \frac{(\mathbb{P}_n - \mathbb{P}) \left([\Phi_{out,\Psi}(\beta^*) - \Phi_{out,\Psi}(\beta)] + [\Phi_{trt,\Psi}(\beta^*) - \Phi_{trt,\Psi}(\beta)] \right)}{\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g - \beta_g^*||_2 + \epsilon_n}$$

with $M = 8 \left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} \right) B + \epsilon_n$ and $\epsilon_n = \frac{1}{n}$.

We can then adapt the following propositions of Blazère, Loubes, and Gamboa (2014):

Proposition 1 *Provided the penalty term λ_n is chosen suitably large enough,*

$$P(\mathcal{A} \cap \mathcal{B}) \geq 1 - \frac{2(C+2)}{(2G_n)^{A^2/2}}$$

for any $A > \sqrt{2}$, where C is a universal constant.

(Proof at the end of this section) In other words, for some suitable values of λ_n and provided

$G_n \to \infty$, the event $\mathcal{A} \cap \mathcal{B}$ happens with probability tending to one, implying the events $\mathcal{A}$ and $\mathcal{B}$ each also have probability tending to one. Propositions 2 and 3 below provide upper bounds for the linear and non-linear parts of the empirical process on the events $\mathcal{A}$ and $\mathcal{A} \cap \mathcal{B}$, each occurring with high probability (by Proposition 1), respectively:

Proposition 2 *On the event $\mathcal{A}$,*

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,l}(\beta^*) - \Phi_{sum,l}(\hat{\beta})) \leq \frac{\lambda_n}{2} \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2.$$

Proof. We have

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,l}(\beta^*) - \Phi_{sum,l}(\hat{\beta})) \\
&= \sum_{g=1}^{G_n} (\hat{\beta}_g - \beta_g^*)^T \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g \right)^T - E\left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g \right)^T \right] \\
&\leq \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} \left\| \hat{\beta}_g - \beta_g^* \right\|_2 \left\| \frac{\hat{\beta}_g^{ls}}{\sqrt{d_g}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i}^g \right)^T - E\left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt}^g \right)^T \right\|_2 .
\end{aligned}$$

The last line follows from the Cauchy-Schwarz inequality, and the proposition follows on the event $\mathcal{A}$. □

Lemma 3 *On the event $\mathcal{A} \cap \mathcal{B}$ we have $\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 \leq M$, where we recall that $M = 8 \left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} \right) B + \epsilon_n$ with $\epsilon_n = \frac{1}{n}$.*

Lemma 3 bounds the difference between the estimated and true coefficients and is proved at the end of this section. The next proposition provides an upper bound for $(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\psi}(\beta^*) - \Phi_{sum,\psi}(\hat{\beta}))$ and directly results from Lemma 3 and definition of $\mathcal{B}$.

Proposition 3 *On the event $\mathcal{A} \cap \mathcal{B}$,*

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\hat{\beta})) \leq \frac{\lambda_n}{2} \left(\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \epsilon_n \right)$$

Lemma 4 *Assume assumptions (H.1-3) are fulfilled. For all $k \in \mathbb{N}^*$, there exists constants $C_{L,B}^{out}$ and $C_{L,B}^{trt}$ (which both only depend on L and B) such that $E(|Y|^k) \leq k!(C_{L,B}^{out})^k$ and $E(|A|^k) \leq k!(C_{L,B}^{trt})^k$.*

L applies to assumption (H.1) and is a uniform bound for the maximum magnitude of the covariates, and B applies to assumption (H.3) and bounds the l_2 norm of the true (grouped) covariates. Lemma 4 provides moment bounds for outcome Y and treatment A and follows from Lemma 3.2 in Blazère et al. (2014) when $\sqrt{d_g}$ in assumption (H.2) in Blazère et al. (2014) is replaced with $\frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}}$.

Theorem 1 requires that the *Group Stabil Condition* be satisfied. We state it here:

Definition 1 *Let $\Sigma = \mathbb{E}[(Z_{out}, Z_{trt})(Z_{out}, Z_{trt})^T]$. Define $H^* = \{g : \beta_g^* \neq 0\}$, the index set of the groups for which the corresponding sub vectors of β^* are non-zero. Let c_0 and $\epsilon > 0$ be given. Then Σ satisfies the Group Stabil condition if there exists $0 < k < 1$ such that*

$$\delta^T \Sigma \delta \ \geq k \sum_{g \in H^*} ||\delta^g||_2^2 - \epsilon$$

for any $\delta \in S(c_0, \epsilon)$, where $S(c_0, \epsilon)$ is called the restricted set and is defined for c_0 and $\epsilon > 0$ as $S(c_0, \epsilon) = \{\delta : \sum_{g \in H^{*c}} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\delta^g||_2 \leq c_0 \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\delta^g||_2 + \epsilon\}$. A Σ which satisfies the Group Stabil Condition is said to be GS(c_0, ϵ, k).

Definition 1 is similar to the Group Stabil Condition proposed in Blazère et al. (2014), the only difference is that $\sqrt{d_g}$ in the restricted set in Blazère et al. (2014) is replaced here by $\frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}}$ in the restricted set, $S(c_0, \epsilon)$. The Group Stabil Condition places a lower bound on the eigenvalues of the variance matrix, with the lower bound depending on the number of non-zero covariate groups. In other words, it restricts the degree of correlation between covariates in the design matrix.

We can now prove Theorem 1 presented in Section 4.2 in the main manuscript:

Proof of Theorem 1:

The proof uses arguments similar to those in Blazère et al. (2014). Using the definition of

$\hat{\beta}$, where we recall from the main manuscript that

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \mathbb{P}_n(\Phi_{sum}(\beta)) + \lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g||_2 \right\},$$

we have

$$\mathbb{P}_n \Phi_{sum}(\hat{\beta}) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g||_2 \leq \mathbb{P}_n \Phi_{sum}(\beta^*) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g^*||_2. \tag{2}$$

Hence we get (adding $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$ to both sides)

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g||_2$$

$$\leq (\mathbb{P}_n - \mathbb{P})(\Phi_{sum}(\beta^*) - \Phi_{sum}(\hat{\beta})) + 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\beta_g^*||_2. \tag{3}$$

From Proposition 2 and 3 and by adding $\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2$ to both sides of the inequality (3) we find, on $\mathcal{A} \cap \mathcal{B}$, that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$$

$$\leq 2\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} (||\hat{\beta}_g - \beta_g^*||_2 + ||\beta_g^*||_2 - ||\hat{\beta}_g||_2) + \frac{\lambda_n}{2} \epsilon_n.$$

If $g \notin H^*$, where we recall from Definition 1 that $H^* = \{g : \beta_g^* \neq 0\}$ (i.e., the index set of the groups for which the corresponding sub vectors of β^* are non-zero), then $||\hat{\beta}_g - \beta_g^*||_2 + ||\beta_g^*||_2 - ||\hat{\beta}_g||_2 = 0$ and otherwise $||\beta_g^*||_2 - ||\hat{\beta}_g||_2 \leq ||\hat{\beta}_g - \beta_g^*||_2$. So the last inequality can be bounded by

$$4\lambda_n \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \frac{\lambda_n}{2} \epsilon_n. \tag{4}$$

By the definition of β^* we have $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) > 0$ and therefore

$$\sum_{g \notin H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 \leq 3 \sum_{g \in H^*} \frac{\sqrt{d_g}}{\hat{\beta}_g^{ls}} ||\hat{\beta}_g - \beta_g^*||_2 + \frac{\epsilon_n}{2},$$

i.e., $\hat{\beta} - \beta^* \in S(3, \frac{\epsilon_n}{2})$. The next proposition provides a lower bound for $\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$.

Proposition 4 *On the event $\mathcal{A} \cap \mathcal{B}$ we have*

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*)) \geq c_n \mathbb{E}\left[(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt}))^2\right]$$

with $c_n := \min_{\left\{|x| \leq L\left(\min_{g \in G_n}\{\hat{\beta}_g^{ls}\} M + B\right)\right\} \cap \Theta_{out}} \left\{\frac{\Psi''_{out}(x)}{2}\right\} + \min_{\left\{|x| \leq L\left(\min_{g \in G_n}\{\hat{\beta}_g^{ls}\} M + B\right)\right\} \cap \Theta_{trt}} \left\{\frac{\Psi''_{trt}(x)}{2}\right\}$

Proof. We have

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{sum}(\beta^*))$$

$$= \mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*)) + \mathbb{P}(\Phi_{trt}(\hat{\beta}) - \Phi_{trt}(\beta^*))$$

$$= \mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*)) + \mathbb{P}(\Phi_{trt}(\hat{\beta}) - \Phi_{trt}(\beta^*)).$$

Recall $\beta = (\alpha, \gamma)^T$ where α are the regression parameters in the outcome model and γ are the regression parameters in the treatment model, and note that

$$\mathbb{P}(\Phi_{out}(\hat{\beta}) - \Phi_{out}(\beta^*))$$

$$= -\mathbb{E}\left[\mathbb{E}\left(Y|Z_{out}\right)\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)\right]$$

$$+\mathbb{E}\left[\psi'_{out}(\alpha^{*T} Z_{out})\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)\right]$$

$$+\mathbb{E}\left[\frac{\psi''_{out}(\tilde{\alpha}^T Z_{out})}{2}\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)^2\right],$$

where $\tilde{\alpha}^T Z_{out}$ is an intermediate point between $\hat{\alpha}^T Z_{out}$ and $\alpha^{*T} Z_{out}$ given by a second order

Taylor expansion of ψ_{out}. Since $\psi'_{out}(\alpha^{*T} Z_{out}) = \mathbb{E}(Y|Z_{out})$ we find

$$\mathbb{P}(\Phi_{out}(\hat{\alpha}) - \Phi_{out}(\alpha^*)) = \mathbb{E}\left[\frac{\psi''_{out}(\tilde{\alpha}^T Z_{out})}{2}\left(\hat{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}\right)^2\right].$$

Besides we have

$$\begin{aligned}
|\tilde{\alpha}^T Z_{out}| &\leq |\tilde{\alpha}^T Z_{out} - \alpha^{*T} Z_{out}| + |\alpha^{*T} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} |\tilde{\alpha}^{gT} Z^g_{out} - \alpha^{*gT} Z^g_{out}| + \sum_{g=1}^{G_n} |\alpha^{*gT} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} |\hat{\alpha}^{gT}_n Z^g_{out} - \alpha^{*gT} Z^g_{out}| + \sum_{g=1}^{G_n} |\alpha^{*gT} Z_{out}| \\
&\leq \sum_{g=1}^{G_n} \left\|\hat{\alpha}_g - \alpha^*_g\right\|_2 \left\|Z^g_{out}\right\|_2 + \sum_{g=1}^{G_n} \left\|\alpha^*_g\right\|_2 \left\|Z^g_{out}\right\|_2,
\end{aligned}$$

where the first inequality and second line follows from the triangle inequality, the third line follows because $\tilde{\alpha}^T Z_{out}$ is between $\hat{\alpha}^T_n Z_{out}$ and $\alpha^{*T} Z_{out}$, and the fourth line follows from Hölder's inequality. Applying (H.1), we find

$$\left\|Z_{out}\right\|_2 \leq L\sqrt{d_g} \to$$

$$|\tilde{\alpha}^T Z_{out}| \leq L\left(\sum_{g=1}^{G_n} \left\|\hat{\alpha}_g - \alpha^*_g\right\|_2 \sqrt{d_g} + \sum_{g=1}^{G_n} \left\|\alpha^*_g\right\|_2 \sqrt{d_g}\right).$$

Then using Lemma 3 and (H.3) we find

$$|\tilde{\alpha}^T Z_{out}| \leq L\left(\min_{g \in G_n}\left\{\hat{\beta}^{ls}_g\right\} M + B\right) \ a.s.$$

Moreover, α^* and $\hat{\alpha}$ belong to Λ_{out}, which is a convex set, so we know $\tilde{\alpha} \in \Lambda_{out}$, and therefore, $\tilde{\alpha}^T Z_{out} \in \Theta_{out}$ a.s. It follows that

$$\mathbb{P}(\Phi_{out}(\hat{\alpha}) - \Phi_{out}(\alpha^*)) \geq c_{1n}\mathbb{E}\left[(\hat{\alpha} Z_{out} - \alpha^* Z_{out})^2\right]$$

where $c_{1n} := \min_{\left\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}M+B\right)\right\}\cap\Theta_{out}} \left\{\frac{\Psi''_{out}(x)}{2}\right\}$.

We can use a similar argument to show

$$\mathbb{P}(\Phi_{trt}(\hat{\gamma}) - \Phi_{trt}(\gamma^*)) \geq c_{2n}\mathbb{E}\left[(\hat{\gamma}Z_{trt} - \gamma^* Z_{trt})^2\right]$$

where $c_{2n} := \min_{\left\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}M+B\right)\right\}\cap\Theta_{trt}} \left\{\frac{\Psi''_{trt}(x)}{2}\right\}$.
Therefore,

$$\mathbb{P}(\Phi_{sum}(\hat{\beta}) - \Phi_{trt}(\beta^*)) \geq (c_{1n}+c_{2n})\mathbb{E}\left[(\hat{\alpha}Z_{out} - \alpha^* Z_{out})^2 + (\hat{\gamma}Z_{trt} - \gamma^* Z_{trt})^2\right]$$

$$\geq (c_{1n}+c_{2n})\mathbb{E}\left[(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt}))^2\right].$$

□

From Proposition 4 and (4) we deduce that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g - \beta_g^*||_2 + c_n\mathbb{E}\left[\left(\hat{\beta}^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt})\right)^2\right]$$

$$\leq 4\lambda_n \sum_{g\in H^*} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g - \beta_g^*||_2 + \frac{\lambda_n}{2}\epsilon_n. \tag{5}$$

Let $\Sigma = \mathbb{E}\left[(Z_{out}, Z_{trt})(Z_{out}, Z_{trt})^T\right]$ be the covariance matrix. We have

$$\mathbb{E}\left[\left(\hat{\beta}_n^T(Z_{out}, Z_{trt}) - \beta^{*T}(Z_{out}, Z_{trt})\right)^2\right] = (\hat{\beta} - \beta^*)^T\Sigma(\hat{\beta} - \beta^*).$$

Because condition $GS(3, \frac{\epsilon_n}{2}, k)$ is satisfied (by assumption) we have

$$c_n(\hat{\beta} - \beta^*)^T\Sigma(\hat{\beta} - \beta^*) \geq c_n k \sum_{g\in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2 - \frac{\epsilon_n}{2}$$

which implies from (5) that

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} \left\| \hat{\beta}_g - \beta_g^* \right\|_2 + c_n k \sum_{g \in H^*} \left\| \hat{\beta}_g - \beta_g^* \right\|_2^2$$

$$\leq 4\lambda_n \sum_{g \in H^*} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} \left\| \hat{\beta}_g - \beta_g^* \right\|_2 + \frac{\lambda_n}{2}\epsilon_n + \frac{\epsilon_n}{2}.$$

Then using the Cauchy-Schwarz inequality on the line above we find

$$\lambda_n \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 + c_n k \sum_{g \in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2$$

$$\leq 4\lambda_n \sqrt{\sum_{g \in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}} \sqrt{\sum_{g \in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2} + (\lambda_n + 1)\frac{\epsilon_n}{2}.$$

Now the fact that $2xy \leq tx^2 + y^2/t$ for all $t > 0$ leads to the following inequality (with $x = 2\lambda_n \sqrt{\sum_{g \in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}}$, $y = \sqrt{\sum_{g \in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2}$, and recalling that $\zeta^* = \sum_{g \in H^*} \frac{d_g}{(\hat{\beta}_g^{ls})^2}$):

$$\lambda_n \sum_{g=1}^{G_n} \sqrt{d_g} ||\hat{\beta}_g - \beta_g^*||_2 + c_n k \sum_{g \in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2$$

$$\leq 4t\lambda_n^2 \zeta^* + \frac{1}{t} \sum_{g \in H^*} ||\hat{\beta}_g - \beta_g^*||_2^2 + (\lambda_n + 1)\frac{\epsilon_n}{2}. \tag{6}$$

Replacing t by $\frac{1}{c_n k}$ in (6) (and dividing by λ_n) we obtain

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq \frac{4}{c_n k}\lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right)\frac{\epsilon_n}{2}.$$

What is more, letting $W_g = \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}$, we have

$$\sum_{g=1}^{G_n} W_g \left\| \left(\hat{\beta}_g - \beta_g^*\right) I(\alpha_g^* \neq 0) \right\|_2 \leq \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2.$$

This yields

$$\sum_{g:\alpha_g^*\neq 0} \left\| \left(\hat{\beta}_g - \beta_g^*\right) \right\|_2 \leq \frac{4}{\min_{g:\alpha_g^*\neq 0}\{W_g\} c_n k} \lambda_n \zeta^* + \left(1 + \frac{1}{\lambda_n}\right) \frac{1}{\min_{g:\alpha_g^*\neq 0}\{W_g\} 2n}.$$

Finally we conclude the proof using Proposition 1.

Proof of Proposition 1:

Let $A > \sqrt{2}$. Recall that we have assumed G_n and n are such that $\frac{\log(2G_n)}{n} \leq 1$. We deduce Proposition 1 from the following two lemmas:

Lemma 5 *Let*

$$\lambda_n \geq \left(8\sqrt{2} A L C_{L,B} \sqrt{\frac{\log(2G_n)}{n}}\right) \vee \left(16 A^2 L C_{L,B} \frac{\log(2G_n)}{n}\right)$$

with $A > 1$. Then

$$\mathbb{P}\{\mathcal{A}\} \geq 1 - 2 d_{max} (2G_n)^{1-A^2}$$

Lemma 6 *Let*

$$\lambda_n \geq 20 A L \left(\max_{(|x|\leq L\kappa_n)\cap\Theta} |\Psi'_{sum}(x)| \right) \sqrt{\frac{2\log(2G_n)}{n}}$$

where $A \geq 1$. Then

$$\mathbb{P}\{\mathcal{B}\} \geq 1 - 2C(2G_n)^{-A^2/2}$$

where we recall $\kappa_n := 17B + \frac{2}{n}$. We can notice that $\mathbb{P}(\mathcal{B})$ tends to 1 as n goes to ∞.

Thus if

$$\lambda_n \geq AKL \left\{ C_{L,B}^* \vee \max_{\{|x|\leq L\kappa_n\}\cap\Theta} |\Psi'_{sum}(x)| \right\} \sqrt{\frac{2\log(2G_n)}{n}}$$

with K chosen such that

$$\lambda_n \geq max(C_1, C_2, C_3)$$

where

$$C_1 := 8\sqrt{2} A L C_{L,B}^* \sqrt{\frac{\log(2G_n)}{n}}$$

$$C_2 := 16A^2LC^*_{L,B}\frac{\log(2G_n)}{n}$$

and

$$C_3 := 20AL\left(\max_{(|x|\leq L\kappa_n)\cap\Theta}|\Psi'_{sum}(x)|\right)\sqrt{\frac{2\log(2G_n)}{n}}$$

then $\mathbb{P}(\mathcal{A}\cap\mathcal{B}) \geq 1-(2d_{max}+2C)(2G_n)^{-A^2/2}$. □

Proof of Lemma 3

The proof of Lemma 3 is based on convexity of the loss function and of the penalty, as in Blazère et al. (2014), where the main idea is similar to the one used by Bühlmann and van de Geer (2011) for the lasso to show consistency of the excess risk. Define $t := \frac{M}{M+\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\hat{\beta}_g-\beta_g^*||_2}$ and $\tilde{\beta} := t\hat{\beta}+(1-t)\beta^*$. By convexity of Φ_{sum} and the L_2 norm, in addition to the fact that $\hat{\beta}$ satisfies (2), we find

$$\mathbb{P}(\Phi_{sum}(\tilde{\beta})-\Phi_{sum}(\beta^*))+2\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\tilde{\beta}_g||_2$$

$$\leq(\mathbb{P}_n-\mathbb{P})(\Phi_{sum}(\beta^*)-\Phi_{sum}(\tilde{\beta}))+2\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g^*||_2.$$

On the event $\mathcal{A}\cap\mathcal{B}$ we have (from Propositions 2 and 3)

$$\mathbb{P}(\Phi_{sum}(\tilde{\beta})-\Phi_{sum}(\beta^*))+2\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\tilde{\beta}_g||_2$$

$$\leq\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\tilde{\beta}_g-\beta_g^*||_2+\lambda_n\frac{\epsilon_n}{2}+2\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g^*||_2.$$

Because $\mathbb{P}(\Phi_{sum}(\tilde{\beta})-\Phi_{sum}(\beta^*))\geq 0$, by adding to both sides of the inequality $2\lambda_n\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g^*||_2$ and by using the triangle inequality, we have

$$\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\tilde{\beta}_g-\beta_g^*||_2\leq\frac{\epsilon_n}{2}+4\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g^*||_2.$$

Therefore, using (H.3), we have

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\tilde{\beta}_g - \beta_g^*||_2 \leq \frac{\epsilon_n}{2} + 4 \min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} B = \frac{M}{2},$$

i.e.,

$$t \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq \frac{M}{2},$$

and then the definition of t leads to

$$\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|} ||\hat{\beta}_g - \beta_g^*||_2 \leq M.$$

□

Proof of Lemma 5:

Proof. We have

$$\mathbb{P}(\mathcal{A}^C) \leq$$

$$\sum_{g=1}^{G_n} \mathbb{P} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i} \right) - E \left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt} \right) \right\} \right\|_2^2 > \frac{\lambda_n^2}{4} d_g \right\} \leq$$

$$\sum_{g=1}^{G_n} \sum_{j=1}^{d_g} \mathbb{P} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \left\{ \left(\frac{Y_i}{sd(Y)} Z_{out,i}^g, A_i Z_{trt,i} \right) - E \left(\frac{Y}{sd(Y)} Z_{out}^g, A Z_{trt} \right) \right\} \right| > \frac{\lambda_n}{2} \right\}. \quad (7)$$

We will define random variables $\{W_{ij}^g\}$ with $j = 1, 2$ (more generally, $j = 1, \ldots, d_g$) and $i = 1, \ldots, n$ such that

$$W_{i1}^g := \frac{Y_i}{sd(Y)} Z_{out,i}^g - \mathbb{E} \left(\frac{Y_i}{sd(Y)} Z_{out}^g \right)$$

and

$$W_{i2}^g := A_i Z_{trt,i}^g - \mathbb{E}(A_i Z_{out}^g)$$

for $i = 1, \ldots, n$. The random variables $\{W_{ij}\}_{i=1,\ldots,n}$ are independent, identically distributed

and centered, and for all $m \geq 2$,

$$\mathbb{E}|W_{i1}^g|^m \leq \sum_{k=0}^{m} \binom{m}{k} \mathbb{E}\left|\frac{Y_i}{sd(Y)} Z_{out,i}\right|^k \left(\mathbb{E}\left|\frac{Y_i}{sd(Y)} Z_{out,i}\right|\right)^{m-k}$$

and

$$\mathbb{E}|W_{i2}^g|^m \leq \sum_{k=0}^{m} \binom{m}{k} \mathbb{E}|A_i Z_{trt,i}|^k (\mathbb{E}|A_i Z_{trt,i}|)^{m-k}.$$

By Jensen's inequality, we obtain

$$\mathbb{E}|W_{i1}^g|^m \leq 2^m \max_{k=1,\ldots,m} \left\{ \mathbb{E}\left|\frac{Y_i}{sd(Y)} Z_{out,i}\right|^k \mathbb{E}\left|\frac{Y_i}{sd(Y)} Z_{out,i}\right|^{m-k} \right\}$$

and

$$\mathbb{E}|W_{i2}^g|^m \leq 2^m \max_{k=1,\ldots,m} \{\mathbb{E}|A_i Z_{trt,i}|^k \mathbb{E}|A_i Z_{trt,i}|^{m-k}\}.$$

For all $k \in \mathbb{N}$, by (H.1) and Lemma 4 we have

$$\mathbb{E}\left|\frac{Y_i}{sd(Y)} Z_{out,i}\right|^k \leq L^k k! (C_{L,B}^{out})^k$$

and

$$\mathbb{E}|A_i Z_{trt,i}|^k \leq L^k k! (C_{L,B}^{trt})^k.$$

Therefore $\mathbb{E}|W_{ij}^g|^m \leq m!(2LC_{L,B}^*)^m$, where $C_{L,B}^* = \max\{C_{L,B}^{out}, C_{L,B}^{trt}\}$. Hence the conditions are satisfied to apply Bernstein's concentration inequality (Bennett, 1962) with $K = 2LC_{L,B}^*$ and $\sigma^2 = 8(LC_{L,B}^*)^2$. Thus we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} W_{ij}^g\right| > \lambda_n/2\right)$$

$$\leq 2\left(\exp\left(\frac{-n\lambda_n}{16LC_{L,B}^*}\right) + \exp\left(\frac{-n\lambda_n^2}{32(2LC_{L,B}^*)^2}\right)\right). \tag{8}$$

Finally, from (7) and (8), we deduce that $\mathbb{P}(\mathcal{A}^c)$ is bounded by

$$2d_{max}G_n\left(\exp\left(\frac{-n\lambda_n}{16LC^*_{L,B}}\right)+\exp\left(\frac{-n\lambda_n^2}{32(2LC^*_{L,B})^2}\right)\right).$$

Therefore if

$$\lambda_n \geq A^2 16LC^*_{L,B}\frac{\log(2G_n)}{n} \vee A8\sqrt{2}LC^*_{L,B}\sqrt{\frac{\log(2G_n)}{n}}$$

with $A>1$ then

$$\mathbb{P}\{\mathcal{A}^c\} \leq 2d_{max}(2G_n)^{1-A^2}.$$

□

Proof of Lemma 6:

Proof. The proof rests on the following Lemma:

Lemma 7 *Let $R>0$ be given. Define*

$$Z_R := \sup_{\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g-\beta_g^*||_2\leq R}\left\{|(\mathbb{P}_n-\mathbb{P})(\Phi_{sum,\Psi}(\beta^*)-\Phi_{sum,\Psi}(\beta))|\right\}.$$

If $A\geq 1$ then

$$\mathbb{P}\left(Z_R \geq A5DLR\sqrt{\frac{2\log(2G_n)}{n}}\right) \leq 2(2G_n)^{-A^2}$$

where $D := \max\left\{\underset{\{|x|\leq L(R+B)\}\cap\Theta}{max}\left\{|\Psi'_{out}(x)+\Psi'_{trt}(x)|\right\}\right\}$.

Proof. Let $R>0$ be given and β satisfy $\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g-\beta_g^*||_2 \leq R$. Then we know $Z_{R,out} := \sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\alpha_g-\alpha_g^*||_2 \leq R$ (and similarly, $Z_{R,trt} := \sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\gamma_g-\gamma_g^*||_2 \leq R$). Notice that if we change X_i by X_i' while keeping the others fixed then $Z_{out,R}$ is modified by at most

$$\frac{2}{n}\left(\underset{g\in\{1,\ldots,G_n\}}{min}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)LR\exp(L(\underset{g\in\{1,\ldots,G_n\}}{min}\left\{|\hat{\beta}_g^{ls}|\right\}R+B)).$$

To see this let

$$\mathbb{P}_n = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j,Y_j}$$

and

$$\mathbb{P}'_n = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j,Y_j} + 1_{X'_j,Y'_j}$$

then we have

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta)) - (\mathbb{P}'_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))$$

$$= (\mathbb{P}_n - \mathbb{P})(\Phi_{out,\Psi_{out}}(\alpha^*) - \Phi_{out,\Psi_{out}}(\alpha)) - (\mathbb{P}'_n - \mathbb{P})(\Phi_{out,\Psi_{out}}(\alpha^*) - \Phi_{out,\Psi_{out}}(\alpha))$$

$$+(\mathbb{P}_n - \mathbb{P})(\Phi_{out,\Psi_{out}}(\gamma^*) - \Phi_{out,\Psi_{out}}(\gamma)) - (\mathbb{P}'_n - \mathbb{P})(\Phi_{out,\Psi_{out}}(\gamma^*) - \Phi_{out,\Psi_{out}}(\gamma))$$

$$= \frac{1}{n}(\Phi_{out,\Psi out}(\alpha^*, Z_{out,i}) - \Phi_{out,\Psi_{out}}(\alpha, Z_{out,i}) - \Phi_{out,\Psi_{out}}(\alpha^*, Z'_{out,i}) + \Phi_{out,\Psi_{out}}(\alpha, Z'_{out,i}))$$

$$+\frac{1}{n}(\Phi_{trt,\Psi trt}(\gamma^*, Z_{trt,i}) - \Phi_{trt,\Psi_{trt}}(\gamma, Z_{trt,i}) - \Phi_{trt,\Psi_{trt}}(\gamma^*, Z'_{trt,i}) + \Phi_{trt,\Psi_{trt}}(\gamma, Z'_{trt,i}))$$

$$\leq \frac{1}{n}|\Psi'(\tilde{\alpha}^T Z_{out,i})||\alpha^{*T} Z_{out,i} - \alpha^T Z_{out,i}| + \frac{1}{n}|\Psi'(\tilde{\alpha}^T Z'_{out,i})||\alpha^{*T} Z'_{out,i} - \alpha^T Z'_{out,i}|$$

$$+\frac{1}{n}|\Psi'(\tilde{\gamma}^T Z_{trt,i})||\gamma^{*T} Z_{trt,i} - \gamma^T Z_{trt,i}| + \frac{1}{n}|\Psi'(\tilde{\gamma}^T Z'_{trt,i})||\gamma^{*T} Z'_{trt,i} - \gamma^T Z'_{trt,i}|$$

where $\tilde{\alpha} Z_{out,i}$ is an intermediate point between $\alpha^T Z_{out,i}$ and $\alpha^{*T} Z_{out,i}$ (using a first order Taylor expansion of the exponential function, as in the proof to Proposition 4). Then, applying (H.1), we find

$$\|Z_{out}\|_2 \leq L\sqrt{d_g} \rightarrow$$

$$|\tilde{\alpha}^T Z_{out}| \leq L\left(\sum_{g=1}^{G_n} \left\|\hat{\alpha}_g - \alpha_g^*\right\|_2 \sqrt{d_g} + \sum_{g=1}^{G_n} \left\|\alpha_g^*\right\|_2 \sqrt{d_g}\right).$$

Then using (H.3) we find

$$|\tilde{\alpha}^T Z_{out}| \leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\} R + B\right).$$

Similarly, it can be shown that

$$|\tilde{\gamma}^T Z_{trt}| \leq L \left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right).$$

Therefore

$$(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta)) - (\mathbb{P}'_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))$$

$$\leq \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{out}} |\Psi'(\tilde{\alpha}^T Z_{out,i})||\alpha^{*T} Z_{out,i} - \alpha^T Z_{out,i}|$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{out}} |\Psi'(\tilde{\alpha}^T Z'_{out,i})||\alpha^{*T} Z'_{out,i} - \alpha^T Z'_{out,i}|$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z_{trt,i})||\gamma^{*T} Z_{trt,i} - \gamma^T Z_{trt,i}|$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z'_{trt,i})||\gamma^{*T} Z'_{trt,i} - \gamma^T Z'_{trt,i}|$$

$$\leq \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{out}} |\Psi'(\tilde{\alpha}^T Z_{out,i})| \sum_{g=1}^{G_n} ||\alpha^* - \alpha||_2 ||Z_{out}^g||_2$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{out}} |\Psi'(\tilde{\alpha}^T Z'_{out,i})| \sum_{g=1}^{G_n} ||\alpha^* - \alpha||_2 ||Z_{out}^g||_2$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z_{trt,i})| \sum_{g=1}^{G_n} ||\gamma^* - \gamma||_2 ||Z_{trt}^g||_2$$

$$+ \frac{1}{n} \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{trt}} |\Psi'(\tilde{\gamma}^T Z'_{trt,i})| \sum_{g=1}^{G_n} ||\gamma^* - \gamma||_2 ||Z_{trt}^g||_2$$

$$\leq \frac{2}{n} \left(\min_{g \in \{1,\ldots,G_n\}} \left\{ |\hat{\beta}_g^{ls}| \right\} / \sqrt{d_g} \right) LR \max_{\{|x| \leq L\left(\min_{g \in G_n} \left\{ \hat{\beta}_g^{ls} \right\} R + B \right)\} \cap \Theta_{out}} \{|\Psi'_{out}(x)|\}$$

$$+\frac{2}{n}\left(\min_{g\in\{1,\dots,G_n\}}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)LR\max_{\{|x|\leq L\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\}\cap\Theta_{trt}}\{|\Psi'_{trt}(x)|\}$$

$$=\frac{4}{n}M_wLRD$$

where $M_w=\left(\min_{g\in\{1,\dots,G_n\}}\left\{|\hat{\beta}_g^{ls}|\right\}/\sqrt{d_g}\right)$.

We can apply McDiarmid's inequality (also called the bounded difference inequality) to Z_R and obtain

$$\mathbb{P}(Z_R-\mathbb{E}Z_R\geq u)\leq\exp\left(-\frac{nu^2}{8M_w^2R^2L^2D^2}\right).$$

Therefore if $\lambda_n\geq ADM_wLR\sqrt{\frac{8\log 2G_n}{n}}$ with $A>0$ then

$$\mathbb{P}(Z_{R,out}-\mathbb{E}Z_{R,out}\geq\lambda_n)\leq(2G_n)^{-A^2}. \tag{9}$$

Now we have to bound the mean $\mathbb{E}Z_R$. To do this, we need the Symmetrization theorem and the contraction principle (see Appendix A of Blazère et al. (2014)), and then let $\epsilon_1,\dots,\epsilon_n$ be a Rademacher sequence independent of $Z_{out,1},\dots,Z_{out,n}$ and $Z_{trt,1},\dots Z_{trt,n}$ and let $S_R:=\{\beta\in\mathbb{R}^p:\sum_{g=1}^{G_n}\frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g-\beta_g^*||_2\leq R\}$. Then by the Symmetrization theorem and the Contraction principle (since ψ is D-lipschitz on the compact set S_R) we have

$$\mathbb{E}Z_R\leq 4D\mathbb{E}\left(\sup_{\beta\in S_R}\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i(\beta^{*T}Z_i-\beta^TZ_i)|\right)$$

$$\leq 4DR\mathbb{E}\left(\max_{g\in\{1,\dots,G_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{|\hat{\beta}_g^{ls}|||Z_i^g||_2}{\sqrt{d_g}}\right|\right),$$

where the last bound follows from Holder's inequality. By applying the theorem below from Blazère et al. (2014) that's a consequence of Hoeffding's inequality, we obtain

$$\mathbb{E}\left(\max_{g\in\{1,\dots,G_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{|\hat{\beta}_g^{ls}|||Z_i^g||_2}{\sqrt{d_g}}\right|\right)\leq M_wL\sqrt{\frac{2\log(2G_n)}{n}}.$$

Theorem. (Blazère et al., 2014) *Let $X_1,\dots,X_n$ be independent random variables on χ and*

$f_1, \ldots, f_n$ real-valued functions on χ which satisfies for all $j = 1, \ldots, p$ and all $i = 1, \ldots, p$ and all $i = 1, \ldots, n$

$$Ef_j(X_i) = 0, |f_j(X_i)| \leq a_{ij}.$$

Then

$$E\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n} f_j(X_i)\right|\right) \leq \sqrt{2\log(2p)}\max_{1\leq j\leq p}\sqrt{\sum_{i=1}^{n} a_{ij}^2}.$$

It follows that

$$\mathbb{E}Z_R \leq 4M_w RLD\sqrt{\frac{2\log(2G_n)}{n}}. \tag{10}$$

□

Thus from (9) and (10) we know that if $A \geq 1$ then

$$\mathbb{P}\left(Z_R \geq ADM_w LR\left(\sqrt{\frac{8\log 2G_n}{n}} + \sqrt{\frac{2\log(2G_n)}{n}}\right)\right) \leq (2G_n)^{-A^2}$$

for all $R > 0$. □

Split up

$$\left\{\beta \in \mathbb{R}^p : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq M\right\},$$

where $M = 8\left(\min_{g\in G_n}\left\{\hat{\beta}_g^{ls}\right\}\right)B + \epsilon_n$, into two sets which are

$$E_1 = \left\{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq \epsilon_n\right\}$$

and

$$E_2 = \left\{\beta : \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq M\right\}$$

$$\subseteq \bigcup_{j=1}^{j_n}\left\{\beta : 2^{j-1}\epsilon_n < \sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 \leq 2^j\epsilon_n\right\}$$

where $j_n := [\log_2(nM)] + 1$ is the smaller integer such that $2j_n\epsilon_n \geq M$. We recall that

$$v_n := \frac{(\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))}{\sum_{g=1}^{G_n} \frac{\sqrt{d_g}}{|\hat{\beta}_g^{ls}|}||\beta_g - \beta_g^*||_2 + \epsilon_n}$$

and to simplify notation let

$$\alpha(\beta, \beta^*) := (\mathbb{P}_n - \mathbb{P})(\Phi_{sum,\Psi}(\beta^*) - \Phi_{sum,\Psi}(\beta))$$

and

$$\Omega(t) := \max\{ \underset{\{|x| \leq L\left(\min_{g \in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\} \cap \Theta_{out}}{max} \{|\Psi'_{out}(x)|\},$$

$$\underset{\{|x| \leq L\left(\min_{g \in G_n}\left\{\hat{\beta}_g^{ls}\right\}R+B\right)\} \cap \Theta_{trt}}{max} \{|\Psi'_{trt}(x)|\}\}.$$

Let $A \geq 1$. Recall that $\kappa_n := 17B + \frac{2}{n} = 2M + B$. On the event E_1,

$$\mathbb{P}\left(\underset{\beta \in E_1}{sup} |v_n(\beta, \beta^*)| \geq A10L\Omega(L\kappa_n)\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq \mathbb{P}\left(\underset{\beta \in E_1}{sup} |\alpha(\beta, \beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq \mathbb{P}\left(\underset{\beta \in E_1}{sup} |\alpha(\beta, \beta^*)| \geq A5L\Omega(L(\epsilon_n + B))\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

given that $2M \geq \epsilon_n$. From Lemma 7 with $R = \epsilon_n$ we deduce

$$\mathbb{P}\left(\underset{\beta \in E_1}{sup} |v_n(\beta, \beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n\sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq 2(2G_n)^{-A^2}. \tag{11}$$

On the event E_2, using the same type of argument as (11) with $R = 2^j\epsilon_n$ (given that

$2M \geq 2^j \epsilon_n$) for all $j = 1, \ldots, j_n$, we find

$$\mathbb{P}\left(\underset{\beta \in E_2}{sup} |v_n(\beta, \beta^*)| \geq A10L\Omega(L\kappa_n)\epsilon_n \sqrt{\frac{2\log(2G_n)}{n}} \right)$$

$$\leq j_n 2(2G_n)^{-A^2}.$$

Finally we have

$$\leq C' 2(2G_n)^{-\frac{A^2}{2}} \tag{12}$$

where C' is a constant (because $j_n = [\log_2(nM)] + 1$ and $n << G_n$) and the result of Lemma 6 follows from (11) and (12) with $C = 1 + C'$. □

Web Appendix C

In Section 5 of the main paper, simulations are presented comparing the performance of GLiDeR, the two-stage model averaged double robust estimator proposed by Cefalu et al. (2016) (which we abbreviate as "MADR"), two standard doubly robust estimators – one using all covariates ("saturated method") and another which selects covariates via "backward selection" (p-stay = 0.05) on the outcome model – and a non-doubly robust method using the adaptive lasso on only the outcome model to select covariates and estimate the average causal effect for ten scenarios ("Scenarios 1–10" presented in Section 5.1 of the main paper) with $p = 10$ covariates (independent and correlated) and sample size $n = 500$. Additional simulation scenarios are presented here exploring the effects of adjusting the number of covariates and sample size in Scenarios 1–9.

Ratio of mean squared error (MSEs) of the doubly robust average causal treatment effect of GLiDeR, backward selection, MADR, and adaptive lasso (denominator) relative to the saturated variable selection method (numerator) for 5 independent covariates and sample size $n = 500$ and for 10 independent covariates and sample size $n = 250$ are shown in Tables 1 and 2, respectively.

Table 1: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 5 covariates, and sample size n = 500 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	MADR MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.09**	1.01	**1.10**	**1.11**
2	**1.08**	1.00	**1.10**	**1.11**
3	**2.77**	1.00	**2.88**	**3.00**
4	**1.00**	1.00	**1.01**	**0.66**
5	**1.63**	**1.04**	**1.62**	**1.64**
6	**16.82**	**0.91**	**18.15**	**16.00**
7	**1.10**	**1.03**	**1.09**	**1.13**
8	**1.21**	**1.03**	**1.21**	**1.29**
9	**1.02**	**1.02**	**1.02**	**1.02**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

Table 2: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 10 covariates, and sample size n = 250 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	MADR MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.13**	**1.03**	**1.15**	**1.18**
2	**1.12**	**1.04**	**1.14**	**1.14**
3	**3.59**	**1.17**	**3.77**	**3.97**
4	**1.04**	**1.02**	**1.05**	**0.90**
5	**1.81**	**0.93**	**1.79**	**1.78**
6	**24.31**	**1.27**	**24.77**	**21.81**
7	**1.19**	**1.06**	**1.16**	**1.21**
8	**1.39**	**1.10**	**1.38**	**1.46**
9	**1.11**	**1.08**	**1.12**	**1.12**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

The same results are shown for 25 independent covariates and sample size $n = 500$ in Table 3, but results were not calculated for MADR due to the relatively large number of covariates. The MSE ratios with 5 covariates (Table 1) are slightly smaller for all methods and scenarios compared to 10 covariates (Table 2 in main manuscript) as the alternative methods (GLiDeR, backward selection, MADR, and adaptive lasso) are generally more efficient than the saturated method when there are more irrelevant variables. This is further seen for GLiDeR, adaptive lasso, and backward selection with 25 covariates (Table 3) as these

Table 3: Ratio of MSE (saturated model MSE / alternative method MSE) for each scenario with independent data, 25 covariates, and sample size n = 500 over 1,000 Monte Carlo datasets.

Scenario	GLiDeR MSE Ratio	Backward Selection MSE Ratio	Adaptive Lasso MSE Ratio
1	**1.13**	**1.12**	**1.15**
2	**1.18**	**1.17**	**1.20**
3	**3.57**	**3.49**	**4.06**
4	**1.06**	**1.06**	**0.98**
5	**1.94**	**1.80**	**1.89**
6	**24.71**	**9.41**	**22.30**
7	**1.23**	**1.11**	**1.25**
8	**1.45**	**1.34**	**1.53**
9	**1.13**	**1.11**	**1.15**

Bold indicates significant difference (5% significance level) between MSEs (testing equality) from the saturated method (full model) vs. the alternative method using the paired t-test.

methods obtain much greater MSE ratios for all scenarios than with 5 and 10 covariates. When the sample size is cut in half ($n = 250$) with 10 covariates (Table 2), the MSE ratios increase in nearly all scenarios for all alternative methods. In other words, the MSE ratios are further away from 1 for all methods and scenarios when the sample size is halved, which seems to suggest the gap in performance between methods is increased with a smaller sample size. As in the main manuscript, the adaptive lasso only considers the outcome model and under-selects an important confounder weakly related to the outcome but strongly associated to the treatment in Scenario 4 and is less efficient than the saturated method even with 25 covariates.

Results are presented below testing generalized cross-validation (GCV) and k-fold cross-validation (kCV) for $k = 2, 5$, and 10 folds on the outcome model for Scenarios 1–10 with 10 independent covariates for Scenarios 1–9 and 100 independent covariates for Scenario 10 and a sample size of 500 for all scenarios. GCV is performed as discussed in Section 3.4 in the main manuscript and k-fold cross-validation chooses the tuning parameter value λ^* as the value λ yielding the smallest average mean squared prediction error across the k test folds. Performance is generally similar for all procedures, but GCV demonstrates the best performance overall at estimating the causal treatment effect in these scenarios, and also has

a computational advantage over kCV (especially for larger k) as it requires the method to be computed only once on the data. Consequently, we recommend using GCV over kCV for model selection.

Table 4: Comparison of tuning parameter selection procedures.

	GCV			2-fold			5-fold			10-fold		
Scenario	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD
1	0.0081	0.00	0.09	0.0081	0.00	0.09	0.0081	0.00	0.09	0.0081	0.00	0.09
2	0.0090	0.00	0.09	0.0089	0.00	0.09	0.0090	0.00	0.10	0.0090	0.00	0.09
3	0.0085	0.00	0.09	0.0091	0.00	0.10	0.0094	0.00	0.10	0.0091	0.00	0.10
4	0.0100	0.00	0.10	0.0100	0.00	0.10	0.0100	0.00	0.10	0.0101	0.01	0.10
5	0.4008	-0.04	0.63	0.4368	-0.05	0.66	0.4371	-0.06	0.66	0.4433	-0.06	0.66
6	0.0958	0.00	0.31	0.1350	-0.01	0.37	0.1057	0.00	0.33	0.1275	0.00	0.36
7	0.7040	-0.05	0.84	0.7175	-0.05	0.85	0.7213	-0.05	0.85	0.7268	-0.05	0.85
8	0.0143	0.01	0.12	0.0141	0.01	0.12	0.0144	0.01	0.12	0.0143	0.01	0.12
9	0.0117	0.00	0.11	0.0117	0.00	0.11	0.0118	0.00	0.11	0.0118	0.00	0.11
10	0.0603	0.00	0.26	0.0638	-0.07	0.24	0.0644	-0.08	0.24	0.0906	-0.15	0.26

Table 5: Covariates selected (average across 1000 samples) by GLiDeR. Though $p = 100$ covariates are considered for Scenario 10, only results for the first two irrelevant variables (X_9 and X_{10}) are shown here.

Scenario	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.07	0.06	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04
2	1.00	0.12	1.00	1.00	0.04	0.04	0.04	0.04	0.03	0.04
3	1.00	1.00	0.07	0.06	0.06	0.02	0.01	0.01	0.01	0.01
4	1.00	1.00	1.00	1.00	1.00	0.05	0.04	0.04	0.04	0.04
5	0.28	0.28	0.80	0.79	0.80	0.06	0.05	0.05	0.06	0.06
6	0.16	0.17	0.08	0.09	0.16	0.02	0.02	0.02	0.02	0.03
7	0.51	0.11	1.00	0.92	0.09	0.06	0.06	0.08	0.06	0.06
8	1.00	1.00	0.59	0.06	0.06	0.05	0.04	0.05	0.05	0.06
9	1.00	0.06	1.00	1.00	0.04	0.05	0.04	0.06	0.05	0.06
10	1.00	0.99	1.00	1.00	0.21	0.20	0.21	0.19	0.01	0.01

Table 6: Bootstrap 95% percentile confidence interval coverage rates by GLiDeR for all scenarios with sample size $n = 500$ and $p = 10$ covariates (except Scenario 10, which has $p = 100$ covariates) across $1,000$ Bootstrap samples. Note that correlated covariates are not considered for Scenario 10.

Scenario	Independent Covariates Coverage Rate	Correlated Covariates Coverage Rate
1	95.1	95.2
2	94.4	94.8
3	94.3	94.1
4	93.5	95.2
5	95.2	94.5
6	94.3	95.5
7	94.4	91.9
8	95.0	93.6
9	94.8	94.7
10	97.1	-

Table 7: Covariates (potential confounders) considered in the lung transplant registry. Each variable is continuous or binary. The mean and standard deviation (if continuous) or frequency and proportion (if binary) of each covariate for BLT and SLT is also shown.

Name	Description	BLT Mean (sd)/ N (%)	SLT Mean (sd)/ N (%)
	Patient characteristics		
AgeP	Age (yrs)	63.6 (2.9)	64.2 (3.1)
BmiP	Body Mass Index	24.5 (7.4)	24.8 (7.4)
DiabP	Diabetes	64 (13%)	41 (9%)
HgtP	Height (cm)	169.8 (9.1)	169.3 (9.3)
O2amt	Oxygen delivered	4.07 (3.07)	3.43 (1.93)
Karn	Karnofsky score > 60	155 (31%)	188 (42%)
LAS	Lung allocation score	35.8 (7.6)	34.0 (3.6)
WhtP	Race (white)	455 (92%)	416 (94%)
SexP	Gender (female)	211 (43%)	208 (47%)
LifeS	Life support ventilator needed	27 (5%)	4 (1%)
Vent	Assisted ventilation needed	68 (14%)	49 (11%)
Vol	Center volume	94.5 (66.5)	71.3 (45.8)
Walk	6 minute walking distance	746.7 (390.7)	719.2 (322.2)
O2rest	Oxygen needed at rest	31 (6%)	36 (8%)
	Donor characteristics		
AgeD	Age (yrs)	36.3 (14.4)	33.7 (14.4)
BlckD	Race (black)	92 (19%)	87 (20%)
BmiD	Body Mass Index	26.0 (5.2)	25.4 (4.9)
Cig	History of cigarette use	74 (15%)	57 (13%)
CMV	Positive cytomegalovirus (CMV) test	302 (61%)	266 (60%)
Cod	Cause of death - traumatic brain injury	224 (45%)	243 (55%)
DiabD	Diabetes	38 (8%)	24 (5%)
ExpD	Expanded donor	65 (13%)	52 (12%)
HgtD	Height (cm)	175.5 (9.4)	175.3 (9.2)
SexD	Gender (female)	146 (30%)	135 (30%)
Dist	Donor to treatment center distance	206 (243.8)	203.3 (246.9)
Po2	Lung PO2	387.2 (148.4)	364.5 (151.3)
	Other characteristics		
Allo	Local or regional (vs. national) allocation	146 (30%)	114 (26%)
HgtR	Height ratio	1.03 (0.05)	1.04 (0.05)
Isch	Ischemic time	5.5 (1.6)	4.0 (1.4)
SexM	Matching gender	125 (25%)	131 (30%)
RaceM	Matching race	330 (67%)	274 (62%)

References

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**, 33–55.

Blazère, M., Loubes, J. M., and Gamboa, F. (2014). Oracle inequalities for a Group Lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory* **60**, 2303–2318.

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory, and Applications. Berlin: Springer.

Cefalu, M., Dominici, F., Arvold, N. D., and Parmigiani, G. (2016). Model averaged double robust estimation. *Biometrics* (ahead of print).

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Statistical Computing* **25**, 1129–1141.