# Omitted variable bias of Lasso-based inference methods: A finite sample analysis*

Kaspar Wüthrich[†] Ying Zhu[‡]

October 21, 2019

**Abstract**

This paper shows in simulations, empirical applications, and theory that Lasso-based inference methods such as post double Lasso and debiased Lasso can exhibit substantial finite sample omitted variable biases in problems with sparse regression coefficients due to Lasso not selecting relevant control variables. This phenomenon can be systematic and occur even when the sample size is large and larger than the number of control variables. On the other hand, we also establish a "robustness" type of result showing that the omitted variable bias remains bounded with high probability even if the prediction errors of the Lasso are unbounded. In empirically relevant settings, our simulations show that OLS with modern standard errors that accommodate many controls can be a viable alternative to Lasso-based inference methods.

**Keywords:** Lasso, post double Lasso, debiased Lasso, OLS, omitted variable bias, limited variability, finite sample analysis

*Alphabetical ordering. Both authors contributed equally to this work. We would like to thank Stéphane Bonhomme, Graham Elliott, Michael Jansson, Ulrich Müller, Andres Santos, and Jeffrey Wooldridge for their comments. We are especially grateful to Yixiao Sun for providing extensive feedback on an earlier draft. This paper was previously circulated as "Behavior of Lasso and Lasso-based inference under limited variability" and "Omitted variable bias of Lasso-based inference methods under limited variability: A finite sample analysis". Ying Zhu acknowledges financial support from a start-up fund from the Department of Economics at UCSD and the Department of Statistics and the Department of Computer Science at Purdue University, West Lafayette.

[†]Department of Economics, University of California, San Diego. Email: kwuthrich@ucsd.edu

[‡]Department of Economics, University of California, San Diego. Email: yiz012@ucsd.edu.

# 1 Introduction

The least absolute shrinkage and selection operator (Lasso), introduced by Tibshirani (1996), has become a standard tool for model selection in high-dimensional problems where the number of covariates ($p$) is larger than or comparable to the sample size ($n$). To make statistical inference on a single parameter of interest (for example, the effect of a treatment or policy), a standard approach is to first use Lasso to select the control variables with nonzero regression coefficients and then to run OLS with the selected controls. However, this approach relies on strong and unrealistic assumptions to ensure that the Lasso selects all the relevant control variables. This has motivated the development of post double Lasso (Belloni et al., 2014b) and debiased Lasso (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014), which have quickly become the most popular methods for making inference in applications with many control variables. The major breakthrough in this literature is that it does not require the coefficients of the relevant controls to be well separated from zero and selection mistakes are shown to have a negligible impact on the asymptotic inference results.

However, the current paper shows that in problems with sparse regression coefficients, underselection of the Lasso can cause post double Lasso and debiased Lasso to exhibit substantial omitted variable biases (OVBs) relative to the standard deviations, even when $n$ is large and larger than $p$ (e.g., when $n = 10000$ and $p = 4000$). We first provide simulation evidence documenting that large OVBs and poor coverage properties of confidence intervals are persistent across a range of empirically relevant settings. Our simulations show that when the non-zero coefficients are small relative to the noise-to-signal ratios, Lasso cannot distinguish these coefficients from zero. As a consequence, Lasso-based inference methods fail to include relevant controls, which results in substantial OVBs (relative to the empirical standard deviation) and undercoverage of confidence intervals.

To explain this phenomenon, we establish theoretical conditions under which it occurs systematically. We develop novel results on the underselection of the Lasso and derive lower bounds on the OVBs of post double Lasso and the debiased Lasso proposed by van de Geer et al. (2014). We choose a finite sample approach which does not rely on asymptotic approximations and allows us to study the OVBs for fixed $n$, $p$, and a fixed number of relevant controls $k$ (even when $\frac{k \log p}{n}$ does not tend to 0). Consistent with our simulation findings, our theoretical analysis shows that the OVBs can be substantial even when $n$ is large and larger than $p$.

While our lower bound results suggest that the OVBs can be substantial relative to the standard deviation even when $\frac{k \log p}{n}$ is "small", surprisingly enough, we can also establish a "robustness" type of result showing that the OVBs of post double Lasso and the debiased Lasso by van de Geer et al. (2014) remain bounded with high probability even if $\frac{k \log p}{n} \to \infty$ and both Lasso steps are inconsistent in terms of the prediction errors.

Let us consider the linear model

$$Y_i = D_i\alpha^* + X_i\beta^* + \eta_i, \tag{1}$$

$$D_i = X_i\gamma^* + v_i. \tag{2}$$

Here $Y_i$ is the outcome, $D_i$ is the treatment variable of interest, and $X_i$ is a $(1 \times p)$-dimensional vector of additional control variables. The goal is to make inference on the treatment effect $\alpha^*$. In the main part of the paper, we focus on post double Lasso and present results for the debiased Lasso in the appendix. Post double Lasso consists of two Lasso selection steps: A Lasso regression of $Y_i$ on $X_i$ and a Lasso regression of $D_i$ on $X_i$. In the third and final step, the estimator of $\alpha^*$, $\tilde{\alpha}$, is obtained from an OLS regression of $Y_i$ on $D_i$ and the union of controls selected in the two Lasso steps.

OVB arises whenever the relevant controls are selected in neither Lasso step. Thus, to study the OVB, one has to understand theoretically when such *double underselection* is likely to occur. This task is difficult because it requires necessary results on the Lasso's inclusion to show that double underselection can occur with high probability and, to our knowledge, no existing result can explain this phenomenon. In this paper, we prove that if the ratios of the absolute values of the non-zero coefficients to the variance of the controls is no greater than half the penalty parameter, Lasso fails to select these controls in both steps with high probability.[1]

[1]Note that the existing Lasso theory requires the penalty parameter to exceed a certain threshold,

This new necessary result is the key ingredient that allows us to derive an explicit lower bound formula for the OVB of $\tilde{\alpha}$. We show that the OVB lower bound can be substantial relative to the standard deviation obtained from the asymptotic distribution in Belloni et al. (2014b) even when $n$ is large and larger than $p$. For example, when $n = 10000$, $p = 4000$, and the control variables are orthogonal to each other, our results imply that the ratio of the OVB lower bound to the standard deviation can be as large as 0.5 when $k = 5$ and 0.84 when $k = 10$. Moreover, keeping $k$ and $\frac{\log p}{n}$ fixed, increasing $n$ will increase the ratio of the OVB lower bound to the standard deviation.

Since OVBs occur when the absolute values of the non-zero coefficients in both Lasso selection steps are small relative to the noise-to-signal ratios, one might ask if the double underselection problem can be mitigated by rescaling the controls. We show that the issue is still present after rescaling the controls and that the OVB lower bound is unaffected. The reason is that any normalization of $X_i$ simply leads to rescaled coefficients and vice versa, while their product stays the same. This result suggests an equivalence between "small" (nonzero) coefficient problems and problems with "limited" variability in the relevant controls. By rescaling the controls, the former can always be recast as the latter and conversely. As a consequence, the OVB lower bound can be substantial relative to the standard deviation even when the omitted relevant controls have small coefficients.

In view of our theoretical results, all else equal, limited variability in the control variables makes it more likely for the Lasso to omit the relevant controls and for the post double Lasso to exhibit substantial OVBs. Limited variability is ubiquitous in applied economic research and there are many instances where it occurs by design. First, limited variability naturally arises from small cells; that is, when there are only a few observations in some of the cells defined by specific covariate values. Small cells are prevalent in flexible specifications that include many two-way interactions and are saturated in at least a subset of covariates (e.g., Belloni et al., 2014a; Chen, 2015; Decker and Schmitz, 2016; Fremstad, 2017; Knaus et al., 2018; Jones et al., 2018;

which depends on the standard deviations of the noise and covariates.

Schmitz and Westphal, 2017).[2] When the covariates are discrete, limited overlap — a major concern in research designs relying on unconfoundedness-type identification assumptions — can be viewed as a small cell problem (e.g., Rothe, 2017). Moreover, categorical covariates, when incorporated through a set of indicator variables, give rise to small cells if some of the categories are sparsely populated. Second, when researchers perform subsample analyses, there are often covariates that exhibit limited variability within subsamples. Third, in times series and "large $T$" panel data applications, persistence in the covariates over time can lead to limited variability. Finally, many empirical settings feature high-dimensional fixed effects, which often suffer from limited variability. Some authors propose to penalize the fixed effects (e.g., Kock and Tang, 2019), while others do not (e.g., Belloni et al., 2016). The results in this paper suggest that penalizing fixed effects can be problematic.

Our results prompt the question of how to make statistical inference (e.g., testing hypotheses about $\alpha^*$ and constructing confidence intervals) in problems where underselection is a concern. In *moderately high-dimensional* settings where $p$ is comparable to but smaller than $n$, OLS constitutes an alternative to Lasso-based inference procedures. We emphasize the moderately high-dimensional regime and OLS because of their relevance in applied economic research.[3] The main challenge for OLS-based inference in settings with many controls is the construction of standard errors, especially when the noise terms exhibit heteroscedasticity and clustering. For instance, Cattaneo et al. (2018b) show that the usual versions of Eicker-White heteroscedasticity robust standard errors are inconsistent under asymptotics where $p$ grows as fast as $n$. Fortunately, several recently developed approaches provide inference procedures for problems with many controls (e.g., Cattaneo et al., 2018b; Jochmans, 2018; Kline et al., 2018; D'Adamo, 2018). In empirically relevant settings, our simulation results show that OLS with the standard errors proposed by Cattaneo et al. (2018b) exhibits a lower bias and better coverage properties than Lasso-based inference methods.

[2] This popular empirical strategy dates back to the original post double Lasso paper by Belloni et al. (2014b)

[3] In our theoretical results, however, $p$ is allowed to exceed $n$.

## 2 Lasso and post double Lasso

### 2.1 The Lasso

Consider the following linear regression model

$$Y_i = X_i\theta^* + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

where $\{Y_i\}_{i=1}^n = Y$ is a $n$-dimensional response vector, $\{X_i\}_{i=1}^n = X$ is a $n \times p$ matrix of covariates with $X_i$ denoting the $i$th row of $X$, $\{\varepsilon_i\}_{i=1}^n = \varepsilon$ is a zero-mean noise vector, and $\theta^*$ is a $p$-dimensional vector of unknown coefficients.

The Lasso estimator of $\theta^*$ is given by

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - X_i\theta)^2 + \lambda \sum_{j=1}^{p} |\theta_j|, \tag{4}$$

where $\lambda$ is the penalization/regularization parameter. For example, if $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$ and $X$ is a fixed design matrix with normalized columns (i.e., $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = b$ for all $j = 1, \ldots, p$), Bickel et al. (2009) set $\lambda = 2\sigma\sqrt{\frac{2b(1+\tau)\log p}{n}}$ (where $\tau > 0$) to establish upper bounds on $\sqrt{\sum_{j=1}^p (\hat{\theta} - \theta^*)^2}$ with high probability guarantee. Wainwright (2009) sets $\lambda$ proportional to $\frac{\sigma}{\phi}\sqrt{\frac{b \log p}{n}}$, where $\phi \in (0, 1]$ is a measure of correlation between the covariates with nonzero coefficients and those with zero coefficients, to establish perfect selection. Both choices can be extended to random designs; for example, each row of $X \in \mathbb{R}^{n \times p}$ is sampled independently from the same normal distribution and $\text{var}(X_{ij}) = b$ for all $j = 1, \ldots, p$.

Other choices of $\lambda$ are available in the literature. For instance, Belloni and Chernozhukov (2013) develop a data-dependent approach and Belloni et al. (2012) and Belloni et al. (2016) propose penalty choices that accommodate heteroscedastic and clustered errors. In the case of nearly orthogonal $X$ (which is typically required to ensure a good performance of the Lasso in fixed designs), these choices of $\lambda$ have a similar scaling as those in Bickel et al. (2009) and Wainwright (2009). Finally, a very popular practical approach for choosing $\lambda$ is cross-validation. However, only few theoretical results exist on the properties of Lasso when $\lambda$ is chosen using cross-validation;

see, for example, Homrighausen and McDonald (2013, 2014) and Chetverikov et al. (2017).

### 2.2 Post double Lasso

The model (1)–(2) implies the following reduced form model for $Y_i$:

$$Y_i = X_i\pi^* + u_i, \tag{5}$$

where $\pi^* = \gamma^*\alpha^* + \beta^*$ and $u_i = \eta_i + \alpha^* v_i$.

The post double Lasso, introduced by Belloni et al. (2014b), essentially exploits the Frisch-Waugh theorem, where the regressions of $Y$ on $X$ and $D$ on $X$ are implemented with Lasso:

$$\hat{\pi} \in \arg\min_{\pi \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - X_i\pi\right)^2 + \lambda_1 \sum_{j=1}^{p} |\pi_j|, \tag{6}$$

$$\hat{\gamma} \in \arg\min_{\gamma \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} \left(D_i - X_i\gamma\right)^2 + \lambda_2 \sum_{j=1}^{p} |\gamma_j|. \tag{7}$$

The final estimator $\tilde{\alpha}$ of $\alpha^*$ is then obtained from an OLS regression of $Y$ on $D$ and the union of selected controls

$$\left(\tilde{\alpha}, \tilde{\beta}\right) \in \arg\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - D_i\alpha - X_i\beta\right)^2 \quad \text{s.t. } \beta_j = 0 \; \forall j \notin \hat{I} = \hat{I}_1 \cup \hat{I}_2, \tag{8}$$

where $\hat{I}_1 = \operatorname{supp}(\hat{\pi})$ and $\hat{I}_2 = \operatorname{supp}(\hat{\gamma})$.

## 3 Evidence on the OVB of post double Lasso

### 3.1 Numerical example

We first illustrate the underselection of the Lasso and its implications for the OVB of post double Lasso using a simple numerical example.

To study the variable selection properties of the Lasso, we simulate data according to the linear model (3), where $X_i \sim \mathcal{N}(0, \sigma_x^2 I_p)$ is independent of $\varepsilon_i \sim \mathcal{N}(0,1)$

and $\{X_i, \varepsilon_i\}_{i=1}^n$ consists of i.i.d. entries. We set $n = 500$, $p = 200$, and consider a sparse setting where $\theta^* = (\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)'$ and $k = 5$. We employ the "standard" theoretical recommendation for the penalty parameter by Bickel et al. (2009).[4]

Figure 1: Average number of selected covariates

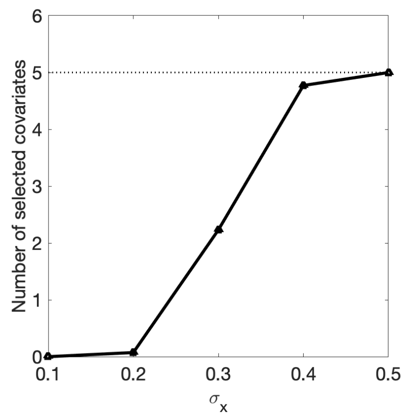



Figure 1 displays the average number of selected covariates as a function of the degree of variability, $\sigma_x$. The variability of the covariates significantly affects the selection performance of the Lasso. The average number of selected covariates is monotonically increasing in $\sigma_x$, ranging from approximately zero when $\sigma_x = 0.1$ up to five when $\sigma_x = 0.5$.

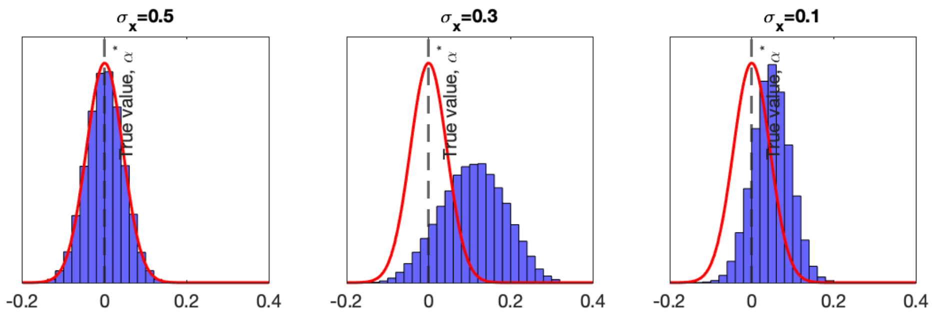
Next, we investigate the implications of the underselection of Lasso for post double Lasso. We simulate data according to the structural model (1)–(2), where $X_i \sim \mathcal{N}(0, \sigma_x^2 I_p)$, $\eta_i \sim \mathcal{N}(0,1)$, and $v_i \sim \mathcal{N}(0,1)$ are independent of each other and $\{X_i, \eta_i, v_i\}_{i=1}^n$ consists of i.i.d. entries. Our object of interest is $\alpha^*$. We set $n = 500$, $p = 200$, $\alpha^* = 0$, and consider a sparse setting where $\beta^* = \gamma^* = (\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)'$ and $k = 5$. We employ the recommendation for the penalty parameter by Bickel et al. (2009).

Figure 2 displays the finite sample distribution of post double Lasso for different values of $\sigma_x$. For comparison, we plot the distribution of the "oracle estimator" of

[4]Specifically, we set $\lambda = 2\sigma\sqrt{\frac{2b(1+\tau)\log p}{n}}$, assuming that $\sigma$ is known. In practice, we first normalize $X_i$ such that $b = 1$, run Lasso using $\lambda = 2\sigma\sqrt{\frac{2(1+\tau)\log p}{n}}$, and then rescale the coefficients.

$\alpha^*$, a regression of $Y_i - X_i\pi^*$ on $D_i - X_i\gamma^*$. When $\sigma_x = 0.5$, the post double Lasso

Figure 2: Finite sample distribution

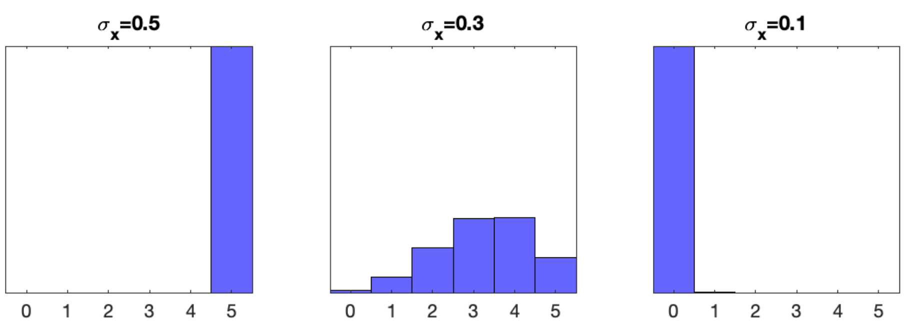



*Notes:* The blue histograms show the finite sample distributions and the red curves show the densities of the oracle estimators.

estimator is approximately unbiased and its distribution is centered at the true value $\alpha^* = 0$. Lowering $\sigma_x$ to 0.3 shifts the distribution to the right and increases the standard deviation. Under very low variability, when $\sigma_x = 0.1$, the distribution is shifted to the right and its shape is similar to the shape when $\sigma_x = 0.5$.

The bias of post double Lasso is caused by the OVB arising from the two Lasso steps not selecting all the important covariates. Figure 3 displays the number of selected important covariates (i.e., the cardinality of $\hat{I}_1 \cup \hat{I}_2$ in (8)). With high prob-

Figure 3: Number of selected relevant controls




ability, all the five important covariates get selected when $\sigma_x = 0.5$. The selection performance deteriorates as $\sigma_x$ decreases, until, with high probability, none of the important covariates get selected when $\sigma_x = 0.1$.

We further note that the shape of the finite sample distribution depends on $\sigma_x$. This distribution is a mixture of the distributions of OLS conditional on the two Lasso steps selecting different combinations of covariates.[5] For $\sigma_x = 0.5$ and $\sigma_x = 0.1$, the finite sample distribution is well-approximated by a normal distribution. The reason is that, when $\sigma_x = 0.5$, the two Lasso steps almost always select all the relevant control variables, whereas none of the relevant controls get selected with high probability when $\sigma_x = 0.1$ (cf. Figure 3). In between these two extreme cases, when $\sigma_x = 0.3$, the finite sample distribution is a mixture of distributions with different means (depending on how many controls get selected), which is skewed and has a larger standard deviation.

The results in this section remain unchanged when, instead of multiplying $X_i$ by $\sigma_x$, we multiply $\beta^*$ and $\gamma^*$ by $\sigma_x$, transforming the limited variability problem into a small coefficients problem. The reason is that $X_i\beta^*$ remains the same in both cases. Thus, one can alternatively interpret and understand the results in this section as showcasing the consequences of small coefficients in settings where the controls exhibit sufficient variability.

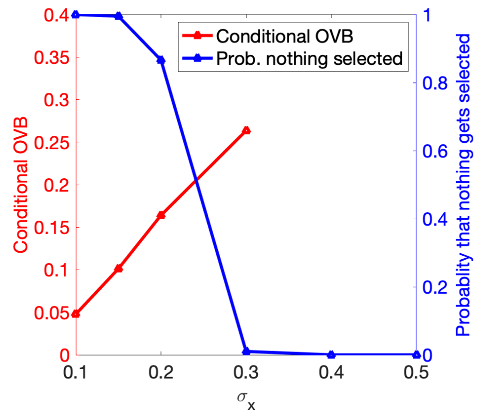
The event where none of the important covariates get selected is of particular interest. Figure 4 displays the probability of this event as a function of $\sigma_x$ as well as the OVB conditional on this event, which is computed as

$$\frac{1}{\sum_{r=1}^{R} S_r} \sum_{r=1}^{R} (\tilde{\alpha}_r - \alpha^*) \cdot S_r.$$

In the above formula, $R$ is the total number of simulation repetitions, $S_r$ is an indicator which is equal to one if nothing gets selected and zero otherwise, and $\tilde{\alpha}_r$ is the estimate of $\alpha^*$ in the $r$th repetition. As the variability in covariates increases from $\sigma_x = 0.1$ to $\sigma_x = 0.5$, the probability that nothing gets selected is decreasing from one to zero. The conditional OVB increases until $\sigma_x = 0.3$ and is not defined for $\sigma_x > 0.3$ because the probability that nothing gets selected is zero in this case.

[5] Note that analytical formulas of these mixture probabilities cannot be derived.

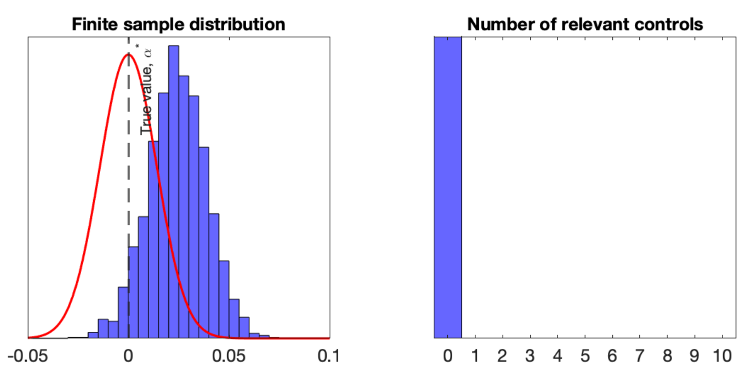
Figure 4: Conditional OVB and probability that nothing gets selected




*Note:* The red (blue) curve is associated with the red (blue) vertical axis.

Importantly, the issue documented here is not a "small sample" phenomenon but persists even in large sample settings. To illustrate, Figure 5 displays the finite sample distributions and the number of selected important covariates for post double Lasso and de-biased Lasso when $n = 5000$, $p = 2000$, $\beta^* = \gamma^* = 0.5 \cdot (\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)'$, $k = 10$, and $\sigma_x = 0.1$. It shows that even in large samples, the finite sample distribution may not be centered at the true value and the bias can be large relative to the standard deviation because none of the important covariates gets selected with high probability.

Figure 5: Performance when $n = 5000$, $p = 2000$, $k = 10$, and $\sigma_x = 0.1$




*Notes:* In the right panel, the blue histogram shows the finite sample distribution and the red curve shows the density of the oracle estimator.

## 3.2 Simulation evidence

Section 3.1 illustrates the implications of underselection due to limited variability based on a simple numerical example and the infeasible penalty choice by Bickel et al. (2009), which assumes that $\sigma^2$ is known. Here we investigate the implications for empirical practice and consider three popular and feasible choices for the penalty parameter $\lambda$: The heteroscedasticity-robust proposal in Belloni et al. (2012) ($\lambda_{\text{BCCH}}$)[6], the penalty parameter with the minimum cross-validated error ($\lambda_{\min}$), and the penalty parameter with the minimum cross-validation error plus one standard deviation ($\lambda_{\text{1se}}$).

The data are simulated according to the structural model

$$Y_i = D_i\alpha^* + X_i\beta^* + \sigma_y(D_i, X_i)\eta_i, \tag{10}$$

$$D_i = X_i\gamma^* + \sigma_d(X_i)v_i, \tag{11}$$

where $X_i \sim \mathcal{N}(0, \sigma_x I_p)$, $\eta_i \sim \mathcal{N}(0,1)$, and $v_i \sim \mathcal{N}(0,1)$ are independent of each other and $\{X_i, \eta_i, v_i\}_{i=1}^n$ consists of i.i.d. entries. The object of interest is $\alpha^*$. We set $n = 500$, $p = 200$, $\alpha^* = 0$, and consider a sparse setting where $\beta^* = \gamma^* = (\underbrace{1, \dots, 1}_{k}, 0, \dots, 0)'$ and $k = 5$.

We study a homoscedastic DGP where $\sigma_y(D_i, X_i) = \sigma_d(X_i) = 1$ and a heteroscedastic DGP where $\sigma_y(D_i, X_i) = \sqrt{\frac{(1+D_i\alpha^*+X_i\beta^*)^2}{\frac{1}{n}\sum_i(1+D_i\alpha^*+X_i\beta^*)^2}}$ and $\sigma_d(X_i) = \sqrt{\frac{(1+X_i\gamma^*)^2}{\frac{1}{n}\sum_i(1+X_i\gamma^*)^2}}$.[7] Appendix D presents additional simulation evidence where we vary $\alpha^*$, the distribution of $X_i$, and the distribution of the error terms $(\eta_i, v_i)$. The results are based on 1,000 repetitions.

---

[6]This approach is based on the following modified Lasso program:

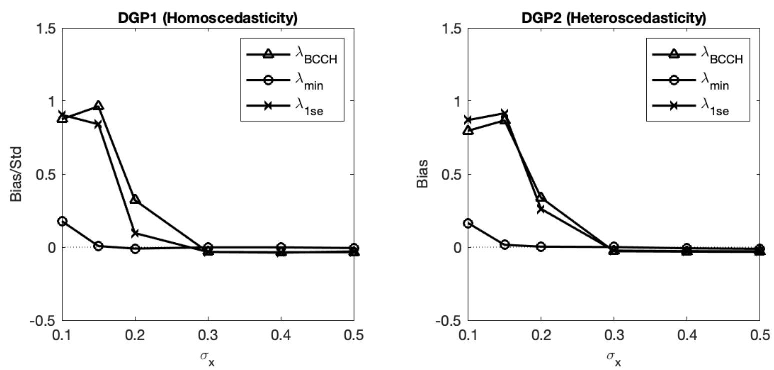
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\theta)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}|\hat{l}_j\theta_j| \tag{9}$$

where $(\hat{l}_1, \dots, \hat{l}_p)$ are penalty loadings obtained using the iterative post Lasso-based algorithm developed in Belloni et al. (2012). Our implementation is based on the Matlab code provided on the authors' webpage: https://voices.uchicago.edu/christianhansen/code-and-data/. We set $\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \varsigma/(2p))$, where $c = 1.1$ and $\varsigma = 0.05$ as recommended by Belloni et al. (2014b).

[7]This multiplicative specification of heteroscedasticity follows the simulation design in Belloni et al. (2014b).

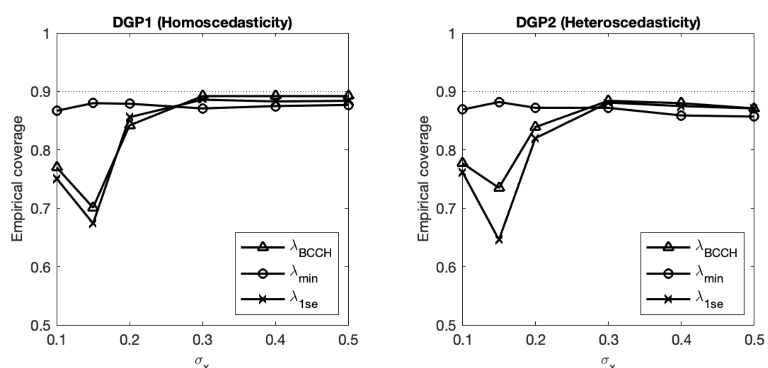
Figure 6 presents evidence on the bias of post double Lasso. To make the results easier to interpret, we report the ratio to the bias and the empirical standard deviation. Under limited variability, the bias of post double Lasso with $\lambda_{\text{BCCH}}$ and $\lambda_{\text{1se}}$

Figure 6: Ratio of bias and standard deviation




is comparable to the standard deviation, whereas the bias with $\lambda_{\text{min}}$ is less than half of the standard deviation. Appendix C shows that these differences are due to the Lasso selecting more relevant controls for all values of $\sigma_x$ with $\lambda_{\text{min}}$ than with $\lambda_{\text{1se}}$ and $\lambda_{\text{BCCH}}$. Under all penalty choices, the ratio of bias to standard deviation decreases to approximately zero as $\sigma_x$ increases to $\sigma_x = 0.3$. Figure 7 displays the coverage rates of 90% confidence intervals and shows that post double Lasso exhibits substantial undercoverage for low values of $\sigma_x$.

Figure 7: Coverage 90% confidence intervals




The additional simulation evidence reported in Appendix D confirms these results and further shows that $\alpha^*$ is an important determinant of the performance of post double Lasso because of its direct effect on the magnitude of the reduced form parameter in (5). Moreover, we show that while choosing $\lambda = \lambda_{\min}$ works well when $\alpha^* = 0$, this choice can yield bad performances when $\alpha^* \neq 0$ (c.f., Figures 17 and 18). As a consequence, based on our simulations, there is no simple recommendation for how to choose $\lambda$ in practice.

The substantive performance differences between the three penalty choices suggest that post double Lasso is sensitive to the penalty level. To further investigate this issue, Figures 8–9 compare the results for $\lambda_{\text{BCCH}}$, $0.5\lambda_{\text{BCCH}}$, and $1.5\lambda_{\text{BCCH}}$. The performance differences are striking. Choosing $\lambda = 0.5\lambda_{\text{BCCH}}$ yields small biases and good coverage properties for all levels of variability considered. By contrast, choosing $\lambda = 1.5\lambda_{\text{BCCH}}$ yields biases that are up to three times larger than the standard deviation and results massive undercoverage.

Figure 8: Ratio of bias and standard deviation: Sensitivity to the penalty choice

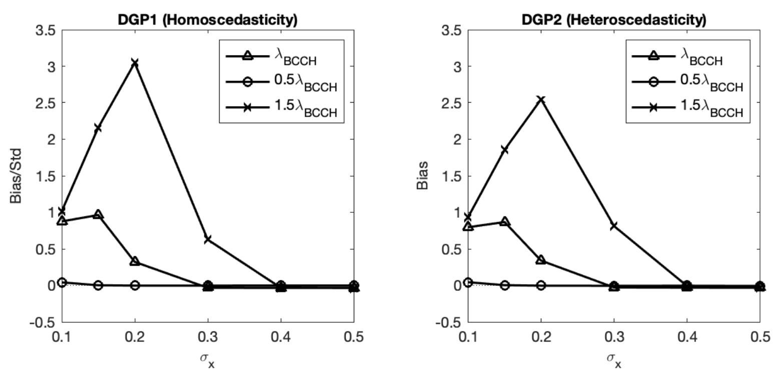
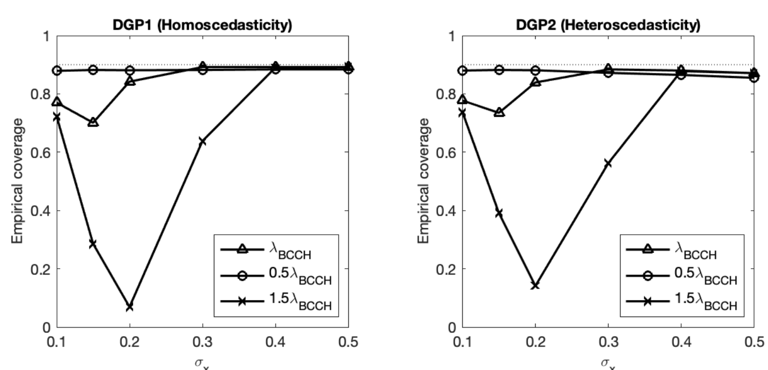



Figure 9: Coverage 90% confidence intervals: Sensitivity to the penalty choice



In sum, our simulation evidence shows that (1) underselection can lead to large biases (compared to the standard deviation) and incorrect inferences and (2) the performance of post double Lasso is very sensitive to the choice of the penalty parameter.

# 4 Theoretical analysis

This section provides a theoretical explanation for the findings in Section 3. We first establish a new necessary result for the Lasso's inclusion and then derive lower bounds on the OVB of post double Lasso. These results do not rely on asymptotic approximations and hold for fixed $n$, $p$, and $k$ (even when $\frac{k\log p}{n}$ does not tend to 0). Most importantly, they also imply that the OVBs can be substantial even if $\frac{k\log p}{n}$ is "small". Surprisingly enough, we can also establish a "robustness" type of result showing that the OVBs remain bounded with high probability even if the prediction errors of the Lasso are unbounded. Throughout this section, we assume a regime where $p$ is comparable to or even much larger than $n$; that is, $p \asymp n$ or $p \gg n$.

For the convenience of the reader, here we collect the notation to be used in the theoretical analyses. Let $1_m$ denote the $m-$dimensional (column) vector of "1"s and $0_m$ is defined similarly. The $\ell_\infty$ matrix norm (maximum absolute row sum) of a matrix $A$ is denoted by $\|A\|_\infty := \max_i \sum_j |a_{ij}|$. For a vector $v \in \mathbb{R}^m$ and a set of indices $T \subseteq \{1, \ldots, m\}$, let $v_T$ denote the sub-vector (with indices in $T$) of $v$. For a matrix $A \in \mathbb{R}^{n\times m}$, let $A_T$ denote the submatrix consisting of the columns with indices in $T$. For a vector $v \in \mathbb{R}^m$, let $\text{sgn}(v) := \{\text{sgn}(v_j)\}_{j=1,\ldots,m}$ denote the sign vector such that $\text{sgn}(v_j) = 1$ if $v_j > 0$, $\text{sgn}(v_j) = -1$ if $v_j < 0$, and $\text{sgn}(v_j) = 0$ if $v_j = 0$.

## 4.1 Model setup

We consider the structural model (1)–(2), which can be written in matrix notation as

$$Y = D\alpha^* + X\beta^* + \eta, \quad (12)$$

$$D = X\gamma^* + v. \quad (13)$$

Following standard practice, we work with centered data, i.e., $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = 0$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0_p$, and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 0$. In matrix notation, the reduced form (5) becomes

$$Y \quad = \quad X\pi^* + u, \tag{14}$$

where $\pi^* = \gamma^*\alpha^* + \beta^*$ and $u = \eta + \alpha^* v$. We make the following assumptions about model (12)–(13).

**Assumption 1.** *The noise terms $\eta$ and $v$ consist of independent entries drawn from $\mathcal{N}\left(0,\, \sigma_\eta^2\right)$ and $\mathcal{N}\left(0,\, \sigma_v^2\right)$, respectively, where $\eta$ and $v$ are independent of each other.*

**Assumption 2.** *The following conditions are satisfied: (i) $\beta^*$ and $\gamma^*$ are exactly sparse with $k(\leq n)$ non-zero coefficients and $K = \left\{j : \beta_j^* \neq 0\right\} = \left\{j : \gamma_j^* \neq 0\right\} \neq \emptyset$; (ii)*

$$\left\| \left(X_{K^c}^T X_K\right)\left(X_K^T X_K\right)^{-1} \right\|_\infty = 1 - \phi \tag{15}$$

*for some $\phi \in (0,\, 1]$, where $K^c$ is the complement of $K$; (iii) $X_j^T X_j = s \neq 0$ for all $j \in K$, and $X_j^T X_j \leq s$ for all $j \in K^c$.*

Part (ii) in Assumption 2 is known as the incoherence condition due to Wainwright (2009). Bühlmann and van de Geer (2011) show that this type of conditions is sufficient and essentially necessary for the Lasso to achieve perfect selection. To provide some intuition for (15), let us consider the simple case where $k = 1$, $X$ is centered (such that $\frac{1}{n}\sum_{i=1}^{n} X = 0_p$), and the columns in $X_{K^c}$ are normalized such that the standard deviations of $X_K$ and $X_j$ (for any $j \in K^c$) are identical; then, $1 - \phi$ is simply the maximum of the absolute (sample) correlations between $X_K$ and each of $X_j$s with $j \in K^c$. If the design $X$ is orthogonal (which is possible if $n \geq p$), then $\phi = 1$.

Assumptions 1 and 2 are generally considered idealistic. Our goal here is to show that, even in these ideal settings, with high probability, the OVBs of post double Lasso can be substantial relative to the standard deviation provided in the existing literature.

## 4.2 Stronger necessary results on the Lasso's inclusion

Post double Lasso exhibits OVB whenever the relevant controls are selected in neither (6) nor (7). To the best of our knowledge, none of the existing results are strong enough to show that, with high probability, Lasso can fail to select the important controls in both steps. Therefore, we first establish a new necessary result for the Lasso's inclusion in Lemma 1. We focus on fixed designs to highlight the essence of the problem; see Appendix E for an extension to random designs.

**Lemma 1.** *[Necessary result on the Lasso's inclusion] In model (3), suppose the $\varepsilon_i$s are independent over $i = 1, \dots, n$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $\sigma \in (0, \infty)$;*[8] *$\theta^*$ is exactly sparse with at most $k$ non-zero coefficients and $K = \left\{j : \theta_j^* \neq 0\right\} \neq \emptyset$. Let Assumption 2(ii)-(iii) hold. We solve the Lasso (4) with $\lambda \geq \frac{2\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$ (where $\tau > 0$). Let $E_1$ denote the event that $\hat{\theta}_j = -sgn\left(\theta_j^*\right)$ for at least one $j \in K$, and $E_2$ denote the event that $\hat{\theta}_l = sgn\left(\theta_l^*\right)$ for at least one $l \in K$ with*

$$|\theta_l^*| \leq \frac{\lambda n}{2s}. \tag{16}$$

*Then, we have*

$$\mathbb{P}\left(E_1 \cap \mathcal{E}\right) = \mathbb{P}\left(E_2 \cap \mathcal{E}\right) = 0. \tag{17}$$

*where $\mathcal{E}$ is defined in (23) of Appendix A.1 and $\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \frac{1}{p^\tau}$.*

*If (16) holds for all $l \in K$, we have*

$$\mathbb{P}\left(\hat{\theta} = 0_p\right) \geq 1 - \frac{1}{p^\tau}. \tag{18}$$

Lemma 1 shows that for large enough $p$, Lasso fails to select any of the relevant control variables with high probability if (16) holds for all $l \in K$ (cf. Figure 5). Suppose that $\lambda = \frac{2\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$, then (16) becomes $\frac{\left|\theta_l^*\right|}{\sigma/\sqrt{\frac{s}{n}}} \leq \phi^{-1}\sqrt{\frac{2(1+\tau)\log p}{n}}$, where the denominator in the left-hand-side is the *noise-to-signal ratio*. This result implies that normalizing $X_j$ to make $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$ for all $j = 1, \dots, p$ does not change Lemma 1. Such normalization simply leads to rescaled coefficients and estimates (by a factor of $\sqrt{\frac{s}{n}}$). In particular, the choice of $\lambda \geq \frac{2\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$ in Lemma 1

[8] The normality of $\varepsilon_i$ can be relaxed without changing the essence of our results.

becomes $\lambda = \lambda_{norm} \geq \frac{2\sigma}{\phi}\sqrt{\frac{2(1+\tau)\log p}{n}}$; also, $\left|\theta_j^*\right| \leq \frac{\lambda n}{2s}$ (where $\lambda \geq \frac{2\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$ without normalization) is replaced by $\sqrt{\frac{s}{n}}\left|\theta_j^*\right| \leq \frac{\lambda_{norm}}{2}$ (where $\lambda_{norm} \geq \frac{2\sigma}{\phi}\sqrt{\frac{2(1+\tau)\log p}{n}}$ with normalization).

**Remark 1.** *Note that (17) implies that $\mathbb{P}\left(\hat{\theta}_l \neq 0\right) \leq \frac{1}{p^\tau}$ for any $l \in K$ subject to (16). In comparison, Wainwright (2009) shows that whenever $\theta_l^* \in \left(\lambda\frac{n}{s}sgn\left(\theta_l^*\right), 0\right)$ or $\theta_l^* \in \left(0, \lambda\frac{n}{s}sgn\left(\theta_l^*\right)\right)$ for some $l \in K$,*

$$\mathbb{P}\left[sgn\left(\hat{\theta}_K\right) = sgn\left(\theta_K^*\right)\right] \leq \frac{1}{2}. \tag{19}$$

*Constant bounds in the form of (19) cannot explain that, with high probability, Lasso fails to select the important covariates in both* (6) *and* (7) *when $p$ is sufficiently large.*

**Remark 2.** *Under the assumptions in Lemma 1, the choices of regularization parameters $\lambda_1$ and $\lambda_2$ coincide with those in Bickel et al. (2009) when $\phi = 1$; e.g., $X$ consists of orthogonal columns, which is possible if $n \geq p$. In the case of $\phi = 1$, the results in Lemma 1 as well as in Sections 4.3 and B.1 hold for any choices of regularization parameters derived from the principle that $\lambda$ should be no smaller than $2\max_{j=1,\ldots,p}\left|\frac{X_j^T\varepsilon}{n}\right|$. These choices constitute what has been used in the vast majority of literature; see, for example, Bickel et al. (2009), Wainwright (2009), Belloni et al. (2012), and Belloni and Chernozhukov (2013).*

## 4.3 Lower bounds on the OVBs

Proposition 1 derives a lower bound formula for the OVB of post double Lasso. We focus on the case where $\alpha^* = 0$ because the conditions required to derive the explicit formula are difficult to interpret when $\alpha^* \neq 0$. The reason is that the error in the reduced form equation (14) involves $\alpha^*$, such that the choice of $\lambda_1$ in (6) depends on the unknown $\alpha^*$. On the other hand, it is possible to provide an easy-to-interpret scaling result (without explicit constants) for the case where $\alpha^* \neq 0$, as we will show in Proposition 2.

**Proposition 1.** *[OVB lower bound] Let Assumptions 1 and 2 hold. Suppose $\lambda_1 = 2\phi^{-1}\sigma_\eta\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$, $\lambda_2 = 2\phi^{-1}\sigma_v\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$; $\beta_j^* = a\frac{\lambda_1 n}{2s}$ and $\gamma_j^* = b\frac{\lambda_2 n}{2s}$ for all*

*$j \in K$ and some constants $a,\, b \in (0,\, 1]$, or $\beta_j^* = a\frac{\lambda_1 n}{2s}$ and $\gamma_j^* = b\frac{\lambda_2 n}{2s}$ for all $j \in K$ and some constants $a,\, b \in [-1,\, 0)$. In terms of $\tilde{\alpha}$ obtained from (8), we have*

$$\mathbb{E}\left(\tilde{\alpha} - \alpha^* | \mathcal{M}\right) \geq \underbrace{\max_{r \in (0,1]} T_1\left(r\right) T_2\left(r\right)}_{:=\underline{OVB}}$$

*where*

$$\begin{aligned} T_1\left(r\right) &= \frac{\left(1+\tau\right) ab\phi^{-2}\sigma_\eta \frac{k \log p}{n}}{4\left(1+\tau\right)\phi^{-2} b^2 \sigma_v \frac{k\log p}{n} + \left(1+r\right)\sigma_v}, \\ T_2\left(r\right) &= 1 - \frac{k}{2}\exp\left(\frac{-b^2\left(1+\tau\right)\log p}{4\phi^2}\right) - \frac{1}{p^\tau} - \exp\left(\frac{-nr^2}{8}\right), \end{aligned}$$

*for any $r \in (0,\, 1]$, and $\mathcal{M}$ is an event with $\mathbb{P}\left(\mathcal{M}\right) \geq 1 - \frac{k}{2}\exp\left(\frac{-b^2(1+\tau)\log p}{4\phi^2}\right) - \frac{2}{p^\tau}$.* [9]

To gauge the magnitude of the OVB, it is instructive to compare the lower bound with $\sigma_{\tilde{\alpha}} = \frac{1}{\sqrt{n}}\frac{\sigma_\eta}{\sigma_v}$, the standard deviation obtained from the asymptotic distribution in Belloni et al. (2014b).[10] Relative to $\sigma_{\tilde{\alpha}}$, the lower bound for the OVB can be quite large even in settings where $n$ is large. In fact, Proposition 1 shows that when $k$ and $\frac{\log p}{n}$ are fixed, increasing $n$ will increase $\frac{\underline{OVB}}{\sigma_{\tilde{\alpha}}}$.

Let us consider the following example: $n = 10000$, $p = 4000$, $\tau = 0.5$, and $X$ consists of orthogonal columns such that $\phi = 1$. Then, $\frac{\underline{OVB}}{\sigma_{\tilde{\alpha}}} = 0.50$, $\mathbb{P}\left(\mathcal{M}\right) \geq 0.86$ when $k = 5$ (so $\frac{k\log p}{n} = 0.004$) and $\frac{\underline{OVB}}{\sigma_{\tilde{\alpha}}} = 0.84$, $\mathbb{P}\left(\mathcal{M}\right) \geq 0.75$ when $k = 10$ (so $\frac{k\log p}{n} = 0.008$). It is important to bear in mind that this number is a theoretical lower bound corresponding to the most favorable case.

Note that the degree of variability in the controls does not enter the OVB lower bound. However, a small $\frac{s}{n}$ makes large non-zero coefficients more difficult to be

[9] Here (and similarly in Proposition 2), we implicitly assume $p$ is sufficiently large such that $1 - k\exp\left(\frac{-b^2(1+\tau)\log p}{4\phi^2}\right) - \frac{4}{p^\tau} > 0$. Indeed, probabilities in the form of "$1 - c^*\exp\left(-c_0^*\log p\right)$" for universal constants $c^*$ and $c_0^*$ are often referred to as the "high probability" guarantees in the literature of (nonasymptotic) high dimensional statistics concerning $p \asymp n$ or $p \gg n$. The event $\mathcal{M}$ is the intersection of $\left\{\hat{I}_1 = \hat{I}_2 = \emptyset\right\}$ and an additional set, both of which occur with high probabilities. The additional set is needed in our analyses for technical reasons. See (39) of Appendix A.2 for the definition of $\mathcal{M}$.

[10] We thank Ulrich Müller for suggesting this comparison.

distinguished from zeros; see (16). In other words, everything else equal, limited variability in the relevant controls makes it more likely for the Lasso to omit the relevant controls and for the post double Lasso to exhibit substantial OVBs. With $n = 10000$, $p = 4000$, $\tau = 0.5$, $\sigma_\eta = 1$ and $\sigma_v = 1$, Lemma 1 says that, with high probability, none of the relevant control variables are selected if $\max_{l\in K} |\theta_l^*| \leq 0.05$ for $\sqrt{\frac{s}{n}} = 1$ and $\max_{l\in K} |\theta_l^*| \leq 0.5$ for $\sqrt{\frac{s}{n}} = 0.1$. Meanwhile, Proposition 1 shows that $\frac{OVB}{\sigma_{\tilde{\alpha}}}$ is the same in both scenarios. In sum, Proposition 1 suggest that (1) limited variability can be recast as a small coefficient problem and vice versa and (2) the OVB can be substantial relative to $\sigma_{\tilde{\alpha}}$ even when the omitted relevant controls have small coefficients.

The next proposition provides the scaling of OVB lower bounds for the case where $\alpha^* \neq 0$. For completeness, we also include the scaling result for the case where $\alpha^* = 0$. This proposition is useful for understanding how OVBs behave approximately as $n$, $p$, and $k$ grow. For functions $f(n)$ and $g(n)$, we write $f(n) \gtrsim g(n)$ to mean that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \lesssim g(n)$ to mean that $f(n) \leq c^{'}g(n)$ for a universal constant $c^{'} \in (0, \infty)$; $f(n) \asymp g(n)$ when $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold simultaneously. As a general rule, $c$ constants denote positive universal constants that are independent of $n$, $p$, $k$, $\sigma_\eta$, $\sigma_v$, $s$, and may change from place to place.

**Proposition 2.** *[Scaling of OVB lower bound] Let Assumptions 1 and 2 hold. Suppose $\phi^{-1} \lesssim 1$ in (15); the regularization parameters in (6) and (7) are chosen in a similar fashion as in Lemma 1 such that $\lambda_1 \asymp \phi^{-1}\sigma_\eta\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$ and $\lambda_2 \asymp \phi^{-1}\sigma_v\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$; for all $j \in K$, $\beta_j^*\gamma_j^* > 0$, $\left|\beta_j^*\right| \leq \frac{\lambda_1 n}{2s}$ and $\left|\gamma_j^*\right| \leq \frac{\lambda_2 n}{2s}$, but $\left|\beta_j^*\right| \asymp \sigma_\eta\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$ and $\left|\gamma_j^*\right| \asymp \sigma_v\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$. Let us consider $\tilde{\alpha}$ obtained from (8).*

*(i) If $\alpha^* = 0$, then there exist positive universal constants $c^\dagger, c_1, c_2, c_3, c^*, c_0^*$ such that*

$$\mathbb{E}\left(\tilde{\alpha} - \alpha^*|\mathcal{M}\right) \geq c^\dagger \frac{\sigma_\eta}{\sigma_v}\min\left\{\frac{k\log p}{n}, 1\right\}\left[1 - c_1 k\exp\left(-c_2\log p\right) - \exp\left(-c_3 n\right)\right], \quad (20)$$

*where $\mathcal{M}$ is an event with $\mathbb{P}\left(\mathcal{M}\right) \geq 1 - c^* k\exp\left(-c_0^*\log p\right)$.*

*(ii) If $\alpha^* \neq 0$, $\alpha^*\gamma_j^* \in (0, -\beta_j^*]$, $\beta_j^* < 0$ for $j \in K$ (or, $\alpha^*\gamma_j^* \in [-\beta_j^*, 0)$, $\beta_j^* > 0$ for $j \in K$), then there exist positive universal constants $c^\dagger, c_4, c_5, c_6, c_1^*, c_2^*$ such that*

$$\mathbb{E}\left(\tilde{\alpha} - \alpha^* | \mathcal{M}\right) \geq c^\dagger \frac{\sigma_\eta}{\sigma_v} \min\left\{\frac{k \log p}{n}, 1\right\}\left[1 - c_4 k \exp\left(-c_5 \log p\right) - \exp\left(-c_6 n\right)\right], \quad (21)$$

*where $\mathbb{P}(\mathcal{M}) \geq 1 - c_1^* k \exp\left(-c_2^* \log p\right)$.*

## 4.4 Upper bounds on the OVBs

So far our lower bound results have suggested that $\frac{\underline{OVB}}{\sigma_{\tilde{\alpha}}}$ can be substantial even when $\frac{k \log p}{n}$ is "small". Interestingly enough, we can establish a "robustness" type of result showing that the OVBs of post double Lasso remain bounded with high probability even if $\frac{k \log p}{n} \to \infty$ and Lasso is inconsistent in the sense that $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(X_i\hat{\beta} - X_i\beta^*\right)^2} \to \infty$, $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(X_i\hat{\gamma} - X_i\gamma^*\right)^2} \to \infty$.

**Proposition 3.** *[Scaling of OVB upper bound] Let Assumptions 1 and 2 hold. Suppose $\phi^{-1} \precsim 1$ in (15); the regularization parameters in (6) and (7) are chosen in a similar fashion as in Lemma 1 such that $\lambda_1 \asymp \phi^{-1}\sigma_\eta\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$ and $\lambda_2 \asymp \phi^{-1}\sigma_v\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$; for all $j \in K$, $\gamma_j^* = \gamma^*$, $\beta_j^*\gamma^* > 0$, $\left|\beta_j^*\right| \leq \frac{\lambda_1 n}{2s}$ and $|\gamma^*| \leq \frac{\lambda_2 n}{2s}$, but $\left|\beta_j^*\right| \asymp \sigma_\eta\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$ and $|\gamma^*| \asymp \sigma_v\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$. Let us consider $\tilde{\alpha}$ obtained from (8). Then for either $\alpha^* = 0$ or $\alpha^* \neq 0$ subject to the conditions in part (ii) of Proposition 2, there exist positive universal constants $c_1^{'}, c_2^{'}, c_3^{'}, c_3^*, c_4^*$ such that*

$$\mathbb{P}\left(\tilde{\alpha} - \alpha^* \leq \overline{OVB} | \mathcal{M}\right) \geq 1 - c_1^{'} k \exp\left(-c_2^{'} \log p\right) - \exp\left(-c_3^{'} n\right)$$

*where $\mathbb{P}(\mathcal{M}) \geq 1 - c_3^* k \exp\left(-c_4^* \log p\right)$ and*

$$\overline{OVB} \asymp \max\left\{\frac{\sigma_\eta}{\sigma_v}\left(\frac{k \log p}{n} \wedge 1\right), \frac{\left(\sqrt{\frac{s}{n}} \vee \sigma_v\right)\sigma_\eta}{\left(\frac{k \log p}{n} \vee 1\right)\sigma_v^2}\right\}.$$

**Remark 3.** *Suppose $\frac{s}{n} \precsim 1$, $\sigma_v \asymp 1$, and $\frac{c^{'} k}{p^{c^{''}}}$ is small for some sufficiently large positive universal constants $c^{'}$ and $c^{''}$. The result above implies that $\overline{OVB} \asymp \frac{\sigma_\eta}{\sigma_v}$ (bounded if $\sigma_\eta \precsim 1$) with high probability even when $\frac{k \log p}{n}$ tends to infinity.*

# 5 OLS as an alternative to Lasso-based inference methods

Our results prompt the question of how to do inference in problems where underselection is a concern, such as, for example, settings where the covariates exhibit limited variability. In many economic applications, $p$ is comparable to but still smaller than $n$. In such moderately high dimensional settings, OLS provides a natural alternative to Lasso-based inference methods such as post double Lasso.

Under classical conditions, OLS is the best linear unbiased estimator. Moreover, under normality and homoscedasticity, OLS admits exact finite sample inference for any fixed $(n, p)$ as long as $\frac{p+1}{n} \leq 1$ (recalling that the number of regression coefficients is $p + 1$ in (1)). While OLS is unbiased, constructing standard errors is challenging when $p$ is large. Under homoscedasticity, existing standard errors are consistent, provided that the estimator of $\sigma_\eta^2$ incorporates a degrees-of-freedom adjustment (e.g., Cattaneo et al., 2018a,b). By contrast, in the heteroscedastic case, Cattaneo et al. (2018b) show that the usual versions of Eicker-White robust standard errors are inconsistent under asymptotics where the number of controls grows as fast as the sample size. This result motivates a very recent literature to develop inference procedures that are valid in settings with many controls (e.g., Cattaneo et al., 2018b; Jochmans, 2018; Kline et al., 2018).

In addition, unlike the Lasso-based inference methods, OLS does not rely on (approximate) sparsity assumptions, which might not be satisfied in applications. In fact, when the parameter space is unrestricted, OLS-based inference exhibits desirable optimality properties (e.g., Armstrong and Kolesar, 2016, Section 4.1).

Figures 10–11 compare the finite sample performance of OLS with heteroscedasticity robust HCK standard errors proposed by Cattaneo et al. (2018b) and post double Lasso. OLS is unbiased (as expected) and exhibits close-to-exact empirical coverage rates, irrespective of the degree of variability in the controls. The additional simulations in Appendix D confirm the excellent performance of OLS with HCK standard errors.

Figure 10: Ratio of bias and standard deviation

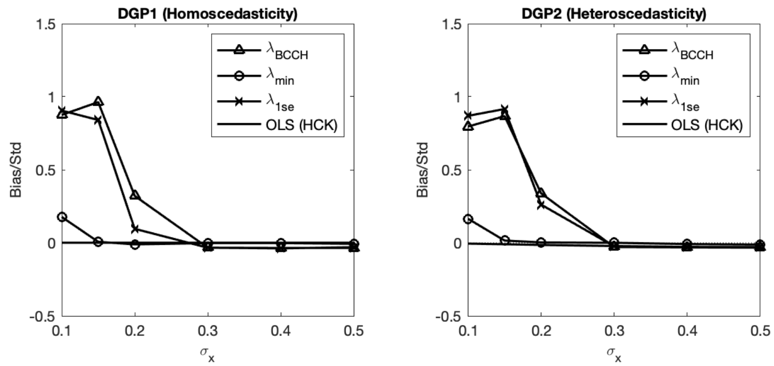



Figure 11: Coverage 90% confidence intervals

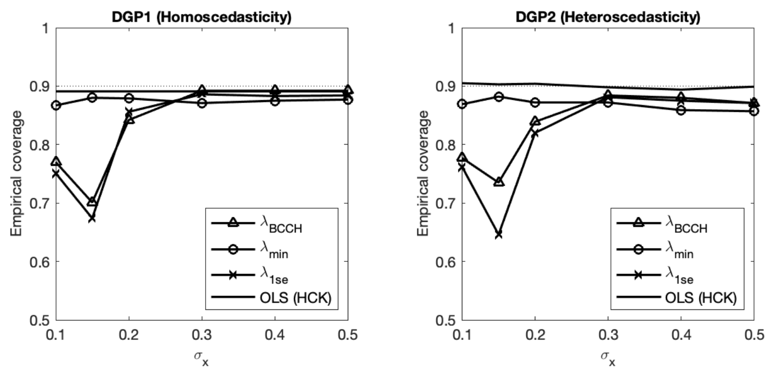



Figure 12: Average length 90% confidence intervals

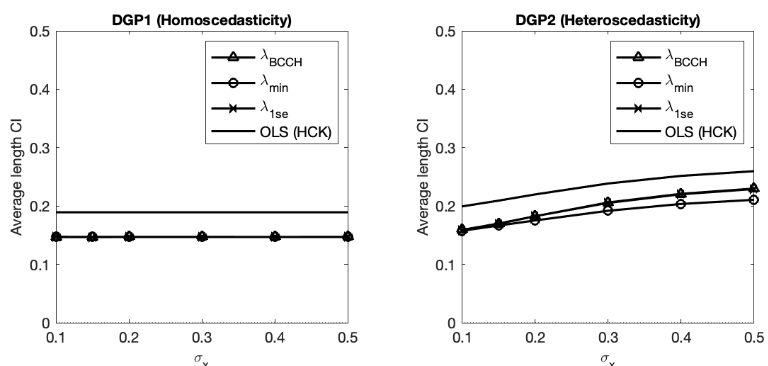



Figure 12 additionally displays the average length of 90% confidence intervals. Under both homoscedasticity and heteroscedasticity, OLS yields somewhat wider confidence intervals than post double Lasso.

In sum, our simulation results suggest that OLS with standard errors that accommodate many controls outperforms post double Lasso in terms of bias and coverage accuracy and thus constitutes a viable alternative in moderately high-dimensional settings.

# 6 Empirical illustration

[To be added.]

# References


Armstrong, T. and Kolesar, M. (2016). Optimal inference in a class of regression models. arXiv 1511.06028v2. 5

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429. 2.1, 3.2, 6, 2, C

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547. 2.1, 2

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50. 1

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. 1, 1, 2, 2.2, 6, 7, 4.3

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605. 1, 2.1

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732. 2.1, 3.1, 3.1, 3.2, 2

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition. 4.1

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):277–301. 5

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361. 1, 5

Chen, D. L. (2015). Can markets stimulate rights? On the alienability of legal claims. *The RAND Journal of Economics*, 46(1):23–65. 1

Chetverikov, D., Chernozhukov, V., and Liao, Z. (2017). On cross-validated lasso. Unpublished Manuscript. 2.1

D'Adamo, R. (2018). Cluster-robust standard errors for linear regression models with many controls. arXiv:1806.07314. 1

Decker, S. and Schmitz, H. (2016). Health shocks and risk aversion. *Journal of Health Economics*, 50:156 – 170. 1

Fremstad, A. (2017). Does craigslist reduce waste? evidence from california and florida. *Ecological Economics*, 132:135 – 143. 1

Homrighausen, D. and McDonald, D. J. (2013). hdm: High-dimensional metrics. *Proceedings of the 30th International Conference on Machine Learning,*, 28. 2.1

Homrighausen, D. and McDonald, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning*, 97(1):65–78. 2.1

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

Jochmans, K. (2018). Heteroskedasticity-robust inference in linear regression models. arXiv:1809.06136. 1, 5

Jones, D., Molitor, D., and Reif, J. (2018). What do workplace wellness programs do? evidence from the illinois workplace wellness study. Working Paper 24229, National Bureau of Economic Research. 1

Kline, P., Saggio, R., and Solvsten, M. (2018). Leave-out estimation of variance components. arXiv:1806.01494. 1, 5

Knaus, M., Lechner, M., and Strittmatter, A. (2018). Heterogeneous employment effects of job search programmes: A machine learning approach. Unpublished Manuscript. 1

Kock, A. B. and Tang, H. (2019). Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory*, 35(2):295?359. 1

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High dimensional ising model selection using $\ell 1$-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319. E.3, E.3

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660. 1

Schmitz, H. and Westphal, M. (2017). Informal care and long-term labor market outcomes. *Journal of Health Economics*, 56:1 – 18. 1

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288. 1

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. 1, B, B.1, B.3, 12

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press. E.1, E.1, 6

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using$\ell_1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202. 2.1, 4.1, 1, 2, A.1, E.3

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. A.1, A.1, E.2

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242. 1, B.1

# Appendix to "Omitted variable bias of Lasso-based inference methods: A finite sample analysis"


**A Proofs for the main results** **2**

A.1 Lemma 1 . . . 2

A.2 Proposition 1 . . . 6

A.3 Proposition 2 . . . 9

A.4 Proposition 3 . . . 10

**B Debiased Lasso** **11**

B.1 Theoretical results . . . 12

B.2 Proof for Propositions 4 and 5 . . . 13

B.3 Simulations evidence . . . 15

**C Lasso selection performance for different feasible penalty choices** **15**

**D Additional simulations** **17**

**E Random design** **20**

E.1 Results . . . 20

E.2 Main proof for Lemma 2 . . . 23

E.3 Additional technical lemmas and proofs . . . 26


## Notation

Here we collect additional notation that is not provided in the main text. The $\ell_q-$norm of a vector $v \in \mathbb{R}^m$ is denoted by $|v|_q$, $1 \leq q \leq \infty$ where $|v|_q := \left(\sum_{i=1}^m |v_i|^q\right)^{1/q}$ when $1 \leq q < \infty$ and $|v|_q := \max_{i=1,\ldots,m} |v_i|$ when $q = \infty$. The support of $v$ is denoted by $\mathrm{supp}(v) := \{j : v_j \neq 0\}$. For a matrix $A \in \mathbb{R}^{n \times m}$, the $\ell_2-$operator norm of $A$ is defined as $\|A\|_2 := \sup_{v \in S^{m-1}} |Av|_2$, where $S^{m-1} = \{v \in \mathbb{R}^m : |v|_2 = 1\}$. For a square matrix $A \in \mathbb{R}^{m \times m}$, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its minimum eigenvalue and maximum eigenvalue, respectively. We denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$.

# A Proofs for the main results

## A.1 Lemma 1

### Preliminary

We will exploit the following Gaussian tail bound:

$$\mathbb{P}\left(\mathcal{Z} \geq t\right) \leq \frac{1}{2} \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

for all $t \geq 0$, where $\mathcal{Z} \sim \mathcal{N}\left(0,\, \sigma^2\right)$. Note that the constant "$\frac{1}{2}$" cannot be improved uniformly.

Given $\lambda \geq \frac{2\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$ where $\tau > 0$ and the tail bound

$$\mathbb{P}\left(\left|\frac{X^T\varepsilon}{n}\right|_\infty \geq t\right) \leq \exp\left(\frac{-nt^2}{2\sigma^2 s/n} + \log p\right) \leq \frac{1}{p^\tau}$$

for $t = \frac{\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2(1+\tau)\log p}{n}}$, we have

$$\lambda \geq 2\left|\frac{X^T\varepsilon}{n}\right|_\infty \tag{22}$$

with probability at least $1 - \frac{1}{p^\tau}$. Let the event

$$\mathcal{E} = \left\{\left|\frac{X^T\varepsilon}{n}\right|_\infty \leq \frac{\sigma}{\phi}\sqrt{\frac{s}{n}}\sqrt{\frac{2\left(1+\tau\right)\log p}{n}}\right\}. \tag{23}$$

Note that $\mathbb{P}\left(\mathcal{E}\right) \geq 1 - \frac{1}{p^\tau}$.

Lemma 1 relies on the following intermediate results.

*(i) On the event $\mathcal{E}$, (4) has a unique optimal solution $\hat{\theta}$ such that $\hat{\theta}_j = 0$ for $j \notin K$.*

*(ii) If $\mathbb{P}\left(\left\{\hat{\theta}_j \neq 0,\, j \in K\right\} \cap \mathcal{E}\right) > 0$, conditioning on $\left\{\hat{\theta}_j \neq 0,\, j \in K\right\} \cap \mathcal{E}$, we must have*

$$\left|\hat{\theta}_j - \theta_j^*\right| \geq \frac{\lambda n}{2s}. \tag{24}$$

Claim (i) above follows from the argument in Wainwright (2019). To show claim (ii), we develop our own proof.

The proof for claim (i) above is based on a construction called Primal-Dual Witness (PDW) method developed by Wainwright (2009). The procedure is described as follows.

1. Set $\hat{\theta}_{K^c} = 0_{p-k}$.

2. Obtain $(\hat{\theta}_K,\, \hat{\delta}_K)$ by solving

$$\hat{\theta}_K \in \arg\min_{\theta_K \in \mathbb{R}^k} \left\{ \underbrace{\frac{1}{2n} |Y - X_K \theta_K|_2^2}_{:=g(\theta_K)} + \lambda |\theta_K|_1 \right\}, \tag{25}$$

   and choosing $\hat{\delta}_K \in \partial |\theta_K|_1$ such that $\nabla g(\theta_K)|_{\theta_K = \hat{\theta}_K} + \lambda \hat{\delta}_K = 0.$[11]

3. Obtain $\hat{\delta}_{K^c}$ by solving

$$\frac{1}{n} X^T (X\hat{\theta} - Y) + \lambda \hat{\delta} = 0, \tag{26}$$

   and check whether or not $\left|\hat{\delta}_{K^c}\right|_\infty < 1$ (the *strict dual feasibility* condition) holds.

Lemma 7.23 from Chapter 7 of Wainwright (2019) shows that, if the PDW construction succeeds, then $\hat{\theta} = (\hat{\theta}_K,\, 0_{p-k})$ is the unique optimal solution of program (4). To show that the PDW construction succeeds on the event $\mathcal{E}$, it suffices to show that $\left|\hat{\delta}_{K^c}\right|_\infty < 1$. The details can be found in Chapter 7.5 of Wainwright (2019). In particular, under the choice of $\lambda$ stated in Lemma 1, we obtain that $\left|\hat{\delta}_{K^c}\right|_\infty < 1$ and hence the PDW construction succeeds conditioning on $\mathcal{E}$ where $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{1}{p^\tau}$.

In summary, conditioning on $\mathcal{E}$, under the choice of $\lambda$ stated in Lemma 1, program (4) has a unique optimal solution $\hat{\theta}$ such that $\hat{\theta}_j = 0$ for $j \notin K$.

We now show (24). By construction, $\hat{\theta} = (\hat{\theta}_K,\, 0_{p-k})$, $\hat{\delta}_K$, and $\hat{\delta}_{K^c}$ satisfy (26) and therefore we obtain

$$\frac{1}{n} X_K^T X_K \left(\hat{\theta}_K - \theta_K^*\right) - \frac{1}{n} X_K^T \varepsilon + \lambda \hat{\delta}_K = 0_k, \tag{27}$$

$$\frac{1}{n} X_{K^c}^T X_K \left(\hat{\theta}_K - \theta_K^*\right) - \frac{1}{n} X_{K^c}^T \varepsilon + \lambda \hat{\delta}_{K^c} = 0_{p-k}. \tag{28}$$

[11] For a convex function $f : \mathbb{R}^p \mapsto \mathbb{R}$, $\delta \in \mathbb{R}^p$ is a subgradient at $\theta$, namely $\delta \in \partial f(\theta)$, if $f(\theta + \triangle) \geq f(\theta) + \langle \delta,\, \triangle \rangle$ for all $\triangle \in \mathbb{R}^p$.

Solving the equations above yields

$$\hat{\theta}_K - \theta^*_K = \left(\frac{X_K^T X_K}{n}\right)^{-1} \frac{X_K^T \varepsilon}{n} - \lambda \left(\frac{X_K^T X_K}{n}\right)^{-1} \hat{\delta}_K. \tag{29}$$

In what follows, we will condition on $\left\{\hat{\theta}_j \neq 0,\, j \in K\right\} \cap \mathcal{E}$ and make use of (22)-(23). Let $\Delta = \frac{X_K^T \varepsilon}{n} - \lambda \hat{\delta}_K$. Note that

$$\left|\hat{\theta}_K - \theta^*_K\right| \geq \left|\left(\frac{X_K^T X_K}{n}\right)^{-1}\right| \left|\left|\lambda \hat{\delta}_K\right| - \left|\frac{X_K^T \varepsilon}{n}\right|\right|, \tag{30}$$

where the inequality uses the fact that $\left(\frac{X_K^T X_K}{n}\right)^{-1}$ is diagonal. In Step 2 of the PDW procedure, $\hat{\delta}_K$ is chosen such that $\left|\hat{\delta}_j\right| = 1$ for any $j \in K$ with $\hat{\theta}_j \neq 0$; we therefore obtain

$$\left|\hat{\theta}_j - \theta^*_j\right| \geq \frac{n}{s} \left| |\lambda| - \left|\frac{X_j^T \varepsilon}{n}\right| \right| \geq \frac{\lambda n}{2s}$$

where the second inequality follows from (22).

**Main proof**

In what follows, we let

$$\begin{aligned}
E_1 &= \left\{\text{sgn}\left(\hat{\theta}_j\right) = -\text{sgn}\left(\theta^*_j\right), \text{ for some } j \in K\right\}, \\
E_2 &= \left\{\text{sgn}\left(\hat{\theta}_j\right) = \text{sgn}\left(\theta^*_j\right), \text{ for some } j \in K \text{ such that (16) holds}\right\}, \\
E_3 &= \left\{\text{sgn}\left(\hat{\theta}_j\right) = \text{sgn}\left(\theta^*_j\right), \text{ for some } j \in K\right\}.
\end{aligned}$$

To show (18) in (iv), recall we have established that conditioning on $\mathcal{E}$, (4) has a unique optimal solution $\hat{\theta}$ such that $\hat{\theta}_j = 0$ for $j \notin K$. Therefore, conditioning on $\mathcal{E}$, the KKT condition for (4) implies

$$\frac{s}{n}\left(\theta^*_j - \hat{\theta}_j\right) = \lambda \text{sgn}\left(\hat{\theta}_j\right) - \frac{X_j^T \varepsilon}{n} \tag{31}$$

for $j \in K$ such that $\hat{\theta}_j \neq 0$.

We first show that $\mathbb{P}\left(E_1 \cap \mathcal{E}\right) = 0$. Suppose $\mathbb{P}\left(E_1 \cap \mathcal{E}\right) > 0$. We may then condition on the event $E_1 \cap \mathcal{E}$. Case (i): $\theta^*_j > 0$ and $\hat{\theta}_j < 0$. Then, the LHS of (31),

$\frac{s}{n}\left(\theta_j^* - \hat{\theta}_j\right) > 0$; consequently, the RHS, $\lambda\text{sgn}\left(\hat{\theta}_j\right) - \frac{X_j^T\varepsilon}{n} = -\lambda - \frac{X_j^T\varepsilon}{n} > 0$. However, given the choice of $\lambda$, conditioning on $\mathcal{E}$, $\lambda \geq 2\left|\frac{X^T\varepsilon}{n}\right|_\infty$ and consequently, $-\lambda - \frac{X_j^T\varepsilon}{n} \leq -\frac{\lambda}{2} < 0$. This leads to a contradiction. Case (ii): $\theta_j^* < 0$ and $\hat{\theta}_j > 0$. Then, the LHS of (31), $\frac{s}{n}\left(\theta_j^* - \hat{\theta}_j\right) < 0$; consequently, the RHS, $\lambda\text{sgn}\left(\hat{\theta}_j\right) - \frac{X_j^T\varepsilon}{n} = \lambda - \frac{X_j^T\varepsilon}{n} < 0$. However, given the choice of $\lambda$, conditioning on $\mathcal{E}$, $\lambda \geq 2\left|\frac{X^T\varepsilon}{n}\right|_\infty$ and consequently, $\lambda - \frac{X_j^T\varepsilon}{n} \geq \frac{\lambda}{2} > 0$. This leads to a contradiction.

It remains to show that $\mathbb{P}(E_2 \cap \mathcal{E}) = 0$. We first establish a useful fact under the assumption that $\mathbb{P}(E_3 \cap \mathcal{E}) > 0$. Let us condition on the event $E_3 \cap \mathcal{E}$. If $\theta_j^* > 0$, we have $\frac{s}{n}\left(\theta_j^* - \hat{\theta}_j\right) = \lambda - \frac{X_j^T\varepsilon}{n} \geq \frac{\lambda}{2} > 0$ (i.e., $\theta_j^* \geq \hat{\theta}_j$); similarly, if $\theta_j^* < 0$, then we have $\frac{s}{n}\left(\theta_j^* - \hat{\theta}_j\right) = -\lambda - \frac{X_j^T\varepsilon}{n} \leq -\frac{\lambda}{2} < 0$ (i.e., $\theta_j^* \leq \hat{\theta}_j$). Putting the pieces together implies that, for $j \in K$ such that $\text{sgn}\left(\hat{\theta}_j\right) = \text{sgn}\left(\theta_j^*\right)$,

$$\left|\theta_j^* - \hat{\theta}_j\right| = \left|\theta_j^*\right| - \left|\hat{\theta}_j\right|. \tag{32}$$

We now show that $\mathbb{P}(E_2 \cap \mathcal{E}) = 0$. Suppose $\mathbb{P}(E_2 \cap \mathcal{E}) > 0$. We may then condition on the event that $E_2 \cap \mathcal{E}$. Because of (16) and (32), we have $\left|\theta_j^* - \hat{\theta}_j\right| < \frac{\lambda n}{2s}$. On the other hand, (24) implies that $\left|\theta_j^* - \hat{\theta}_j\right| \geq \frac{\lambda n}{2s}$. We have arrived at a contradiction. Consequently, we must have $\mathbb{P}(E_2 \cap \mathcal{E}) = 0$.

In summary, we have shown that $\mathbb{P}(E_1 \cap \mathcal{E}) = 0$ and $\mathbb{P}(E_2 \cap \mathcal{E}) = 0$. Claim (i) in "Preliminary" implies that $\mathbb{P}(E_4|\mathcal{E}) = 0$ where $E_4$ denotes the event that $\hat{\theta}_j \neq 0$ for some $j \notin K$. Therefore, on $\mathcal{E}$, none of the events $E_1$, $E_2$ and $E_4$ can happen. This fact implies that, if (16) is satisfied for all $l \in K$, we must have

$$\mathbb{P}\left(\hat{\theta}_K = 0_k\right) \geq 1 - \mathbb{P}(\mathcal{E}^c) \geq 1 - \frac{1}{p^\tau}.$$

## A.2 Proposition 1

We first show the case where $\beta_j^* = a\frac{\lambda_1 n}{2s}$ and $\gamma_j^* = b\frac{\lambda_2 n}{2s}$ for all $j \in K$ and some constants $a,\, b \in (0,\, 1]$. Let the events

$$\begin{aligned} \mathcal{E}_{t_1} &= \left\{ \frac{X_j^T v}{n} \geq -t_1,\, t_1 > 0,\, \forall j \in K \right\}, \quad (33) \\ \mathcal{E}_{t_2}^{'} &= \left\{ \frac{1}{n}\sum_{i=1}^{n} v_i^2 \leq \sigma_v^2 + t_2,\; t_2 \in (0,\, \sigma_v^2] \right\}. \end{aligned}$$

Note that $\mathbb{P}\left(\mathcal{E}_{t_1}\right) = 1 - \mathbb{P}\left(\tilde{\mathcal{E}}_{t_1}\right)$ where $\tilde{\mathcal{E}}_{t_1}$ is the event that $\frac{X_j^T v}{n} \leq -t_1$ for some $j \in K$. By tail bounds for Gaussian and Chi-Square variables, we have

$$\begin{aligned} \mathbb{P}\left(\tilde{\mathcal{E}}_{t_1}\right) &\leq \frac{k}{2}\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right), \quad (34) \\ \mathbb{P}\left(\mathcal{E}_{t_2}^{'}\right) &\geq 1 - \exp\left(\frac{-nt_2^2}{8\sigma_v^4}\right). \end{aligned}$$

From (34), we obtain

$$\mathbb{P}\left(\mathcal{E}_{t_1}\right) \geq 1 - \frac{k}{2}\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right). \quad (35)$$

In the following proof, we exploit the bound

$$\begin{aligned} \mathbb{P}\left(\mathcal{E}_{t_2}^{'} | \hat{I}_2 = \emptyset,\, \mathcal{E}_{t_1}\right) &\geq \mathbb{P}\left(\mathcal{E}_{t_2}^{'} \cap \mathcal{E}_{t_1} \cap \left\{\hat{I}_2 = \emptyset\right\}\right) \\ &\geq \mathbb{P}\left(\mathcal{E}_{t_1}\right) + \mathbb{P}\left(\mathcal{E}_{t_2}^{'}\right) + \mathbb{P}\left(\hat{I}_2 = \emptyset\right) - 2 \\ &\geq 1 - \frac{1}{p^\tau} - \frac{k}{2}\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right) - \exp\left(\frac{-nt_2^2}{8\sigma_v^4}\right) \quad (36) \end{aligned}$$

where the third inequality follows from Lemma 1, which implies $\hat{I}_2 = \emptyset$ with probability at least $1 - \frac{1}{p^\tau}$. Note that $\mathbb{P}\left(\mathcal{E}_{t_1} \cap \left\{\hat{I}_2 = \emptyset\right\}\right) \geq \mathbb{P}\left(\mathcal{E}_{t_1}\right) + \mathbb{P}\left(\hat{I}_2 = \emptyset\right) - 1 \geq 1 - \frac{1}{p^\tau} - \frac{k}{2}\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right)$, which is a "high probability" guarantee for sufficiently large $p$ and $t_1$. Thus, working with $\mathbb{P}\left(\mathcal{E}_{t_2}^{'} | \hat{I}_2 = \emptyset,\, \mathcal{E}_{t_1}\right)$ is sensible under an appropriate choice of $t_1$ (as we will see below).

We first bound $\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*$. Note that

$$\begin{aligned}\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^* &= \left(\frac{D^T D}{n}\right)^{-1}\left[\frac{1}{n}\left(X_K\gamma_K^* + v\right)^T X_K\beta_K^*\right] \\ &= \left(\frac{D^T D}{n}\right)^{-1}\left[\frac{1}{n}\gamma_K^{*T}X_K^T X_K\beta_K^* + \frac{1}{n}v^T X_K\beta_K^*\right] \\ &= \frac{\frac{s}{n}\gamma_K^{*T}\beta_K^* + \frac{1}{n}v^T X_K\beta_K^*}{\frac{1}{n}\left(X_K\gamma_K^* + v\right)^T\left(X_K\gamma_K^* + v\right)}.\end{aligned}$$

Note that $\frac{s}{n}\gamma_K^{*T}\beta_K^* = 2\left(1+\tau\right)ab\phi^{-2}\sigma_\eta\sigma_v\frac{k\log p}{n}$. Moreover, applying (36) with $t_1 = b\frac{\lambda_2}{4}$ and $t_2 = r\sigma_v^2$ yields

$$\frac{s}{n}\gamma_K^{*T}\beta_K^* + \frac{1}{n}v^T X_K\beta_K^* \geq \left(1+\tau\right)ab\phi^{-2}\sigma_\eta\sigma_v\frac{k\log p}{n} \tag{37}$$

as well as

$$\begin{aligned}\frac{1}{n}\left(X_K\gamma_K^* + v\right)^T\left(X_K\gamma_K^* + v\right) &\leq \frac{s}{n}\gamma_K^{*T}\gamma_K^* + \frac{2}{n}v^T X_K\gamma_K^* + \frac{1}{n}v^T v \\ &\leq 4\left(1+\tau\right)\phi^{-2}b^2\sigma_v^2\frac{k\log p}{n} + \sigma_v^2 + r\sigma_v^2\end{aligned}$$

with probability at least

$$1 - \frac{k}{2}\exp\left(\frac{-b^2\left(1+\tau\right)\log p}{4\phi^2}\right) - \frac{1}{p^\tau} - \exp\left(\frac{-nr^2}{8}\right) := T_2\left(r\right).$$

Conditioning on $\mathcal{E}_{t_1}\cap\left\{\hat{I}_2 = \emptyset\right\}$ with $t_1 = t^* = b\frac{\lambda_2}{4}$, putting the pieces together yields

$$\frac{D^T X_K}{D^T D}\beta_K^* \geq \frac{\left(1+\tau\right)ab\phi^{-2}\sigma_\eta\frac{k\log p}{n}}{4\left(1+\tau\right)\phi^{-2}b^2\sigma_v\frac{k\log p}{n} + \sigma_v + r\sigma_v} := T_1\left(r\right), \tag{38}$$

with probability at least $T_2\left(r\right)$. That is,

$$\mathbb{P}\left(\frac{D^T X_K}{D^T D}\beta_K^* \geq T_1\left(r\right)|\hat{I}_2 = \emptyset,\, \mathcal{E}_{t^*}\right) \geq T_2\left(r\right).$$

When $\alpha^* = 0$ in (12), the reduced form coefficients $\pi^*$ in (14) coincide with $\beta^*$ and $u$ coincides with $\eta$. Given the conditions on $X$, $\eta$, $v$, $\beta_K^*$ and $\gamma_K^*$, we can then apply (18) in Lemma 1 and the fact $\mathbb{P}\left(\hat{I}_1 = \hat{I}_2 = \emptyset\right) \geq \mathbb{P}\left(\hat{I}_1 = \emptyset\right) + \mathbb{P}\left(\hat{I}_2 = \emptyset\right) - 1$ to show that $E = \left\{\hat{I}_1 = \hat{I}_2 = \emptyset\right\}$ with probability at least $1 - \frac{2}{p^\tau}$. Note that with the choice

$t_1 = t^* = b\frac{\lambda_2}{4}$, $\mathbb{P}\left(E \cap \mathcal{E}_{t^*}\right) \geq \mathbb{P}\left(E\right) + \mathbb{P}\left(\mathcal{E}_{t^*}\right) - 1 \geq 1 - \frac{k}{2}\exp\left(\frac{-b^2(1+\tau)\log p}{4\phi^2}\right) - \frac{2}{p^{\tau}}$, which is a "high probability" guarantee given sufficiently large $p$. Therefore, it is sensible to work with $\mathbb{E}\left(\tilde{\alpha} - \alpha^* | \mathcal{M}\right)$ where

$$\mathcal{M} = E \cap \mathcal{E}_{t^*}. \tag{39}$$

Given $E$, (8) becomes

$$\tilde{\alpha} \in \arg\min_{\alpha \in \mathbb{R}} \frac{1}{2n} \left|Y - D\alpha\right|_2^2, \qquad \text{while } \tilde{\beta} = 0_p. \tag{40}$$

As a result, we obtain $\mathbb{E}\left(\tilde{\alpha} - \alpha^* | \mathcal{M}\right) = \mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^* | \mathcal{M}\right) + \mathbb{E}\left(\frac{\frac{1}{n}D^T \eta}{\frac{1}{n}D^T D} | \mathcal{M}\right)$ and

$$\begin{aligned}
\mathbb{E}\left(\frac{\frac{1}{n}D^T\eta}{\frac{1}{n}D^T D}|\mathcal{M}\right) &= \frac{1}{\mathbb{P}\left(\mathcal{M}\right)}\mathbb{E}\left[\frac{\frac{1}{n}D^T\eta}{\frac{1}{n}D^T D}1_{\mathcal{M}}\left(D,\eta\right)\right] \\
&= \frac{1}{\mathbb{P}\left(\mathcal{M}\right)}\mathbb{E}_D\left\{\mathbb{E}_\eta\left[\frac{\frac{1}{n}D^T\eta}{\frac{1}{n}D^T D}1_{\mathcal{M}}\left(D,\eta\right)|D\right]\right\} \\
&= \frac{1}{\mathbb{P}\left(\mathcal{M}\right)}\mathbb{E}_D\left\{\frac{\frac{1}{n}\sum_{i=1}^n D_i \mathbb{E}_\eta\left[\eta_i 1_{\mathcal{M}}\left(D,\eta\right)|D\right]}{\frac{1}{n}D^T D}\right\} \\
&= 0
\end{aligned} \tag{41}$$

where $1_{\mathcal{M}}\left(D,\eta\right) = 1\left\{(v,\eta):\ \hat{I}_1 = \hat{I}_2 = \emptyset,\ \frac{X_j^T v}{n} \geq -t^*\,\forall j \in K\right\}$ (recall $X$ is a fixed design); the last line follows from $\frac{1}{n}\sum_{i=1}^n D_i = 0$, the distributional identicalness of $(\eta_i)_{i=1}^n$ and that $\mathbb{E}_\eta\left[\eta_i 1_{\mathcal{M}}\left(D,\eta\right)|D\right]$ is a constant over $i$s.

Given $\alpha^* = 0$, it remains to bound $\mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*|\mathcal{M}\right) = \mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*|\hat{I}_2 = \emptyset,\ \mathcal{E}_{t^*}\right)$. Note that conditioning on $\mathcal{E}_{t^*}$, $\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*$ is positive by (37). Applying a Markov inequality yields

$$\mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*|\hat{I}_2 = \emptyset,\ \mathcal{E}_{t^*}\right) \geq T_1\left(r\right)\mathbb{P}\left(\frac{D^T X_K}{D^T D}\beta_K^* \geq T_1\left(r\right)|\hat{I}_2 = \emptyset,\ \mathcal{E}_{t^*}\right) \geq T_1\left(r\right)T_2\left(r\right).$$

Combining the result above with (41) and maximizing over $r \in (0,\,1]$ gives the claim.

For the case where $\beta_j^* = a\frac{\lambda_1 n}{2s}$ and $\gamma_j^* = b\frac{\lambda_2 n}{2s}$ for all $j \in K$ and some constants $a,\, b \in [-1,\, 0)$, the argument is similar except that we replace (33) with

$$\mathcal{E}_{t_1} = \left\{\frac{X_j^T v}{n} \leq t_1,\ t_1 > 0,\ \forall j \in K\right\}. \tag{42}$$

Note that a similar argument for (35) implies that the event above also holds with probability at least $1 - \frac{k}{2}\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right)$.

## A.3 Proposition 2

Part (i) of Proposition 2 follows immediately from the proof for Proposition 1. It remains to establish part (ii) where $\alpha^* \neq 0$, $\alpha^*\gamma_j^* \in (0, -\beta_j^*]$, $\beta_j^* < 0$ for all $j \in K$ (or, $\alpha^*\gamma_j^* \in [-\beta_j^*, 0)$, $\beta_j^* > 0$ for all $j \in K$). Because of these conditions, we have

$$\left|\pi_j^*\right| = \left|\beta_j^* + \alpha^*\gamma_j^*\right| < \left|\beta_j^*\right| \quad \forall j \in K.$$

Note that $|\alpha^*| \le \max_{j\in K} \frac{\left|\beta_j^*\right|}{\left|\gamma_j^*\right|} \asymp \frac{\sigma_\eta}{\sigma_v}$ and

$$\begin{aligned}
\left|\frac{X^T u}{n}\right|_\infty &= \left|\frac{X^T\left(\eta + \alpha^* v\right)}{n}\right|_\infty \\
&\le \left|\frac{X^T\eta}{n}\right|_\infty + \left|\frac{\alpha^* X^T v}{n}\right|_\infty \\
&\precsim \sigma_\eta\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}} + \frac{\sigma_\eta}{\sigma_v}\sigma_v\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}} \\
&\precsim \phi^{-1}\sigma_\eta\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}
\end{aligned}$$

with probability at least $1 - c_1^{'}\exp\left(-c_2^{'}\log p\right)$. The fact above justifies the choice of $\lambda_1$ stated in 2. We can then apply Lemma 1 (18) to show that $\hat{I}_1 = \emptyset$ with probability at least $1 - c_5\exp\left(-c_6\log p\right)$. Furthermore, under the condition on $\gamma_K^*$, (18) in Lemma 1 implies that $\hat{I}_2 = \emptyset$ with probability at least $1 - c_0\exp\left(-c_0^{'}\log p\right)$. Therefore, we have

$$\mathbb{P}\left(\hat{I}_1 = \hat{I}_2 = \emptyset\right) \ge \mathbb{P}\left(\hat{I}_1 = \emptyset\right) + \mathbb{P}\left(\hat{I}_2 = \emptyset\right) - 1 \ge 1 - c_1^{''}\exp\left(-c_2^{''}\log p\right).$$

Given $u = \eta + \alpha^* v$, when $\alpha^* \neq 0$, the event $\left\{\hat{I}_1 = \emptyset\right\}$ is not independent of $D$, so $\mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^* | E, \mathcal{E}_{t^*}\right) \neq \mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^* | \hat{I}_2 = \emptyset, \mathcal{E}_{t^*}\right)$ (recalling $E = \left\{\hat{I}_1 = \hat{I}_2 = \emptyset\right\}$). Instead of (33) (or, (42)) and (36), we apply

$$\begin{aligned}
\mathcal{E}_{t_1} &= \left\{\left|\frac{X_K^T v}{n}\right|_\infty \le t_1\right\}, \qquad (43) \\
\mathbb{P}\left(\mathcal{E}_{t_1}\right) &\ge 1 - k\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right),
\end{aligned}$$

and

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{t_2}^{'}|E,\,\mathcal{E}_{t_1}\right) &\geq \mathbb{P}\left(\mathcal{E}_{t_2}^{'}\cap\mathcal{E}_{t_1}\cap E\right)\\
&\geq \mathbb{P}\left(\mathcal{E}_{t_1}\right)+\mathbb{P}\left(\mathcal{E}_{t_2}^{'}\right)+\mathbb{P}\left(E\right)-2\\
&\geq 1-c_1^{''}\exp\left(-c_2^{''}\log p\right)-k\exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right)\\
&\quad -\exp\left(\frac{-nt_2^2}{8\sigma_v^4}\right),\qquad \text{for any } t_2\in(0,\,\sigma_v^2],
\end{aligned}
$$

along with the inequalities

$$
\begin{aligned}
\frac{\frac{1}{n}D^TX_K}{\frac{1}{n}D^TD}\beta_K^* &\geq \frac{\frac{s}{n}\gamma_K^{*T}\beta_K^*-|\beta_K^*|_1\left|\frac{1}{n}X_K^Tv\right|_\infty}{\frac{1}{n}\left(X_K\gamma_K^*+v\right)^T\left(X_K\gamma_K^*+v\right)},\\
\frac{1}{n}\left(X_K\gamma_K^*+v\right)^T\left(X_K\gamma_K^*+v\right) &\leq \frac{s}{n}\gamma_K^{*T}\gamma_K^*+|\gamma_K^*|_1\left|\frac{2}{n}X_K^Tv\right|_\infty+\frac{1}{n}v^Tv.
\end{aligned}
$$

The rest of the proof follows from the argument for Proposition 1 and the bounds above.

## A.4 Proposition 3

We make use of the following bound on Chi-Square variables:

$$
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n v_i^2-\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n v_i^2\right)\leq -\sigma_v^2r\right]\leq \exp\left(\frac{-nr^2}{16}\right) \tag{44}
$$

for all $r\geq 0$. On the event $\mathcal{M}=E\cap\mathcal{E}_{t_1}$ where $t_1=t^*=\frac{\gamma^*s}{4n}$ in (43), choosing $r=\frac{1}{2}$ in (44) yields

$$
\begin{aligned}
\frac{\frac{1}{n}D^TX_K}{\frac{1}{n}D^TD}\beta_K^* &\leq \frac{\frac{s}{n}\gamma_K^{*T}\beta_K^*+|\beta_K^*|_1\,t^*}{\frac{s}{n}\gamma_K^{*T}\gamma_K^*-2\,|\gamma_K^*|_1\,t^*+\frac{1}{2}\sigma_v^2}\\
&\leq \frac{c_3\sigma_\eta\sigma_v\frac{k\log p}{n}}{c_1\frac{k\log p}{n}\sigma_v^2+c_2\sigma_v^2}\\
&\leq c_4\frac{\sigma_\eta}{\sigma_v}\left(\frac{k\log p}{n}\wedge 1\right)
\end{aligned}
$$

with probability at least $1-c_5k\exp\left(-c_6\log p\right)-\exp\left(\frac{-n}{64}\right)$.

We can also show that

$$
\begin{aligned}
& \mathbb{P}\left(\frac{1}{n} D^T \eta \leq t \mid \mathcal{M}\right) \\
\geq\ & \mathbb{P}\left(\left\{(\eta, v):\ \frac{1}{n} D^T \eta \leq t\right\} \cap \mathcal{M}\right) \\
\geq\ & \mathbb{P}\left(\frac{1}{n} D^T \eta \leq t\right) + \mathbb{P}(\mathcal{M}) - 1 \\
\geq\ & 1 - c_5 k \exp\left(-c_6 \log p\right) - \exp\left(-c_7 n\left(\frac{t^2}{\left(\frac{s}{n} \vee \sigma_v^2\right) \sigma_\eta^2} \wedge \frac{t}{\left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta}\right)\right).
\end{aligned}
$$

Choosing $t \asymp \left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta$ above yields $\frac{1}{n} D^T \eta \precsim \left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta$ with probability at least $1 - c_5 k \exp\left(-c_6 \log p\right) - \exp\left(-c_8 n\right)$, conditioning on $\mathcal{M}$. We have already shown that, conditioning on $\mathcal{M}$, $\frac{1}{n} D^T D \succsim \left(\frac{k \log p}{n} \vee 1\right) \sigma_v^2$ with probability at least $1 - c_5 k \exp\left(-c_6 \log p\right) - \exp\left(\frac{-n}{64}\right)$. As a consequence,

$$
\begin{aligned}
& \mathbb{P}\left\{\frac{\frac{1}{n} D^T \eta}{\frac{1}{n} D^T D} \precsim \frac{\left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta}{\left(\frac{k \log p}{n} \vee 1\right) \sigma_v^2} \mid \mathcal{M}\right\} \\
\geq\ & \mathbb{P}\left\{\frac{1}{n} D^T \eta \precsim \left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta \textit{ and } \frac{1}{n} D^T D \succsim \left(\frac{k \log p}{n} \vee 1\right) \sigma_v^2 \mid \mathcal{M}\right\} \\
\geq\ & \mathbb{P}\left\{\frac{1}{n} D^T \eta \precsim \left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta \mid \mathcal{M}\right\} + \mathbb{P}\left\{\frac{1}{n} D^T D \succsim \left(\frac{k \log p}{n} \vee 1\right) \sigma_v^2 \mid \mathcal{M}\right\} - 1 \\
\geq\ & 1 - c_9 k \exp\left(-c_{10} \log p\right) - \exp\left(-c_{11} n\right).
\end{aligned}
$$

Putting the pieces above together yields

$$
\mathbb{P}\left(\tilde{\alpha} - \alpha^* \leq \overline{OVB} \mid \mathcal{M}\right) \geq 1 - c_1' k \exp\left(-c_2' \log p\right) - \exp\left(-c_3' n\right)
$$

where $\overline{OVB} \asymp \max\left\{\frac{\sigma_\eta}{\sigma_v}\left(\frac{k \log p}{n} \wedge 1\right), \frac{\left(\sqrt{\frac{s}{n}} \vee \sigma_v\right) \sigma_\eta}{\left(\frac{k \log p}{n} \vee 1\right) \sigma_v^2}\right\}$.

# B Debiased Lasso

In this section, we present theoretical and simulation results on the OVB of the debiased Lasso proposed by van de Geer et al. (2014).

## B.1 Theoretical results

The idea of debiased Lasso is to start with an initial Lasso estimate $\hat{\theta}=\left(\hat{\alpha},\,\hat{\beta}\right)$ of $\theta^{*}=(\alpha^{*},\,\beta^{*})$ in equation (1), where

$$\left(\hat{\alpha},\,\hat{\beta}\right)\in\arg\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^{p}}\frac{1}{2n}\left|Y-D\alpha-X\beta\right|_{2}^{2}+\lambda_{1}\left(\left|\alpha\right|+\left|\beta\right|_{1}\right).\tag{45}$$

Given the initial Lasso estimate $\hat{\alpha}$, the debiased Lasso adds a correction term to $\hat{\alpha}$ to reduce the bias introduced by regularization. In particular, the debiased Lasso takes the form

$$\tilde{\alpha}=\hat{\alpha}+\frac{\hat{\Omega}_{1}}{n}\sum_{i=1}^{n}Z_{i}^{T}\left(Y_{i}-Z_{i}\hat{\theta}\right),\tag{46}$$

where $Z_{i}=(D_{i},\,X_{i})$ and $\hat{\Omega}_{1}$ is the first row of $\hat{\Omega}$, which is an approximate inverse of $\frac{1}{n}Z^{T}Z$, $Z=\{Z_{i}\}_{i=1}^{n}$. Several different strategies have been proposed for constructing the approximate inverse $\hat{\Omega}$; see, for example, Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014). We will focus on the widely used method proposed by van de Geer et al. (2014), which sets

$$\begin{aligned}\hat{\Omega}_{1}&:=&\hat{\tau}_{1}^{-2}\begin{pmatrix}1&-\hat{\gamma}_{1}&\cdots&-\hat{\gamma}_{p}\end{pmatrix},\\ \hat{\tau}_{1}^{2}&:=&\frac{1}{n}\left|D-X\hat{\gamma}\right|_{2}^{2}+\lambda_{2}\left|\hat{\gamma}\right|_{1},\end{aligned}$$

where $\hat{\gamma}$ is defined in (7).

**Proposition 4.** *[Scaling of OVB lower bound for debiased Lasso] Let Assumptions 1 hold. Suppose: with probability at least $1-\kappa$, $\left\|\left(Z_{-K}^{T}X_{K}\right)\left(X_{K}^{T}X_{K}\right)^{-1}\right\|_{\infty}\leq1-\frac{\phi}{2}$ for some $\phi\in(0,\,1]$ such that $\phi^{-1}\precsim1$, where $Z_{-K}$ denotes the columns in $Z=(D,\,X)$ excluding $X_{K}$; the regularization parameters in (7) and (45) are chosen in a similar fashion as in Lemma 1 such that $\lambda_{1}\asymp\phi^{-1}\left(\sqrt{\frac{s}{n}}\vee\sigma_{v}\right)\sigma_{\eta}\sqrt{\frac{\log p}{n}}$ and $\lambda_{2}\asymp\phi^{-1}\sigma_{v}\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$; for all $j\in K$, $\beta_{j}^{*}\gamma_{j}^{*}>0$, $\left|\beta_{j}^{*}\right|\leq\frac{\lambda_{1}n}{2s}$ and $\left|\gamma_{j}^{*}\right|\leq\frac{\lambda_{2}n}{2s}$, but $\left|\beta_{j}^{*}\right|\asymp\left[\sqrt{\frac{n}{s}}\vee\frac{n\sigma_{v}}{s}\right]\sigma_{\eta}\sqrt{\frac{\log p}{n}}$ and $\left|\gamma_{j}^{*}\right|\asymp\sigma_{v}\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$. Let us consider $\tilde{\alpha}$ obtained from (46). If $\alpha^{*}=0$, then there exist positive universal constants $c^{\dagger},c_{7},c_{8},c_{9},c_{3}^{*},c_{4}^{*}$ such that*

$$\mathbb{E}\left(\tilde{\alpha}|\mathcal{M}^{'}\right)\geq c^{\dagger}\frac{\sigma_{\eta}}{\sigma_{v}}\left(\frac{k\log p}{n}\wedge1\right)\left[1-2\kappa-c_{7}k\exp\left(-c_{8}\log p\right)-\exp\left(-c_{9}n\right)\right],$$

*where $\mathcal{M}^{'}$ is an event with $\mathbb{P}\left(\mathcal{M}^{'}\right)\geq1-2\kappa-c_{3}^{*}k\exp\left(-c_{4}^{*}\log p\right)$.*

**Proposition 5.** *[Scaling of OVB upper bound for debiased Lasso] Let Assumptions 1 hold. Suppose: with probability at least $1-\kappa$, $\left\|\left(Z_{-K}^{T}X_{K}\right)\left(X_{K}^{T}X_{K}\right)^{-1}\right\|_{\infty}\leq 1-\frac{\phi}{2}$ for some $\phi\in(0,\,1]$ such that $\phi^{-1}\precsim 1$, where $Z_{-K}$ denotes the columns in $Z=(D,\,X)$ excluding $X_K$; the regularization parameters in (7) and (45) are chosen in a similar fashion as in Lemma 1 such that $\lambda_1\asymp\phi^{-1}\left(\sqrt{\frac{s}{n}}\vee\sigma_v\right)\sigma_\eta\sqrt{\frac{\log p}{n}}$ and $\lambda_2\asymp\phi^{-1}\sigma_v\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$; for all $j\in K$, $\gamma_j^*=\gamma^*$, $\beta_j^*\gamma^*>0$, $\left|\beta_j^*\right|\leq\frac{\lambda_1 n}{2s}$ and $|\gamma^*|\leq\frac{\lambda_2 n}{2s}$, but $\left|\beta_j^*\right|\asymp\left[\sqrt{\frac{n}{s}}\vee\frac{n\sigma_v}{s}\right]\sigma_\eta\sqrt{\frac{\log p}{n}}$ and $|\gamma^*|\asymp\sigma_v\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$. Let us consider $\tilde{\alpha}$ obtained from (46). If $\alpha^*=0$, Then there exist positive universal constants $c_1^{''},c_2^{''},c_3^{''},c_5^{''},c_6^{''}$ such that*

$$\mathbb{P}\left(\tilde{\alpha}-\alpha^*\leq\overline{OVB}|\mathcal{M}^{'}\right)\geq 1-c_1^{''}k\exp\left(-c_2^{''}\log p\right)-\exp\left(-c_3^{''}n\right)$$

*where $\mathbb{P}\left(\mathcal{M}^{'}\right)\geq 1-c_5^{''}k\exp\left(-c_6^{''}\log p\right)$ and*

$$\overline{OVB}\asymp\max\left\{\frac{\sigma_\eta}{\sigma_v}\left(\frac{k\log p}{n}\wedge 1\right),\,\frac{\left(\sqrt{\frac{s}{n}}\vee\sigma_v\right)\sigma_\eta}{\left(\frac{k\log p}{n}\vee 1\right)\sigma_v^2}\right\}.$$

**Remark 4.** *One can show that a population version of the mutual incoherence condition, i.e., $\left\|\left[\mathbb{E}\left(Z_{-K}^{T}\right)X_{K}\right]\left(X_{K}^{T}X_{K}\right)^{-1}\right\|_{\infty}=1-\phi$, implies $\left\|\left(Z_{-K}^{T}X_{K}\right)\left(X_{K}^{T}X_{K}\right)^{-1}\right\|_{\infty}\leq 1-\frac{\phi}{2}$ with high probability (that is, $\kappa$ is small and vanishes polynomially in $p$). For example, we can apply (78) in Lemma 5 with slight notational changes.*

**Remark 5.** *Like in Proposition 2, the event $\mathcal{M}^{'}$ in Proposition 4 is the intersection of $\left\{\hat{\beta}=\hat{\gamma}=0_p\right\}$ and an additional set, both of which occur with high probabilities. The additional set is needed in our analyses for technical reasons. See Appendix B.2 for details.*

## B.2 Proof for Propositions 4 and 5

Under the conditions in Proposition 4, (18) in Lemma 1 implies that $\hat{\gamma}=0_p$ with probability at least $1-c_0\exp\left(-c_0^{'}\log p\right)$. Conditioning on this event, $\hat{\Omega}_1=\left(\frac{1}{n}D^TD\right)^{-1}e_1$ where $e_1=\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}$. If $\alpha^*=0$, under the conditions in Proposition 4, we show that $\hat{\beta}=0_p$ with probability at least $1-2\kappa-c_5\exp\left(-c_6\log p\right)$. To achieve this goal, we slightly modify the argument for (18) in Lemma 1 by replacing (23) with $\mathcal{E}=\mathcal{E}_1\cap\mathcal{E}_2$, where

$$\begin{aligned}\mathcal{E}_1 &= \left\{\left|\frac{Z^T\eta}{n}\right|_\infty \precsim \phi^{-1}\left(\sqrt{\frac{s}{n}} \vee \sigma_v\right)\sigma_\eta\sqrt{\frac{\log p}{n}}\right\},\\ \mathcal{E}_2 &= \left\{\left\|\left(Z_{-K}^T X_K\right)\left(X_K^T X_K\right)^{-1}\right\|_\infty \leq 1-\frac{\phi}{2}\right\},\end{aligned}$$

and $Z_{-K}$ denotes the columns in $Z$ excluding $X_K$. Note that by (65), $\mathbb{P}(\mathcal{E}_1) \geq 1 - c_1^{'}\exp\left(-c_2^{'}\log p\right)$ and therefore, $\mathbb{P}(\mathcal{E}) \geq 1-\kappa-c_1^{'}\exp\left(-c_2^{'}\log p\right)$. We then follow the argument used in the proof for Lemma 1 to show $\mathbb{P}(E_1\cap\mathcal{E}) = 0$ and $\mathbb{P}(E_2\cap\mathcal{E}) = 0$, where

$$\begin{aligned}E_1 &= \left\{\mathrm{sgn}\left(\hat{\beta}_j\right) = -\mathrm{sgn}\left(\beta_j^*\right),\text{ for some } j\in K\right\},\\ E_2 &= \left\{\mathrm{sgn}\left(\hat{\beta}_j\right) = \mathrm{sgn}\left(\beta_j^*\right),\text{ for some } j\in K\right\}.\end{aligned}$$

Moreover, conditioning on $\mathcal{E}$, $\hat{\beta}_{K^c} = 0_{p-k}$. Putting these facts together yield the claim that $\hat{\beta} = 0_p$ with probability at least $1-2\kappa-c_5\exp(-c_6\log p)$.

Letting $E = \left\{\hat{\beta} = \hat{\gamma} = 0_p\right\}$ with $\mathbb{P}(E) \geq 1-2\kappa-c_1\exp(-c_2\log p)$ and recalling the event $\mathcal{E}_{t^*}$ in the proof for 1(ii), we can then show

$$\begin{aligned}\mathbb{E}\left(\tilde{\alpha}-\alpha^*|\mathcal{M}^{'}\right) &= \frac{1}{n}\mathbb{E}\left(\hat{\Omega}_1 Z^T\eta|\mathcal{M}^{'}\right) + \mathbb{E}\left[\frac{D^T X_K}{D^T D}\left(\beta_K^*-\hat{\beta}_K\right)|\mathcal{M}^{'}\right]\\ &= \mathbb{E}\left(\frac{\frac{1}{n}D^T\eta}{\frac{1}{n}D^T D}|\mathcal{M}^{'}\right) + \mathbb{E}\left(\frac{D^T X_K}{D^T D}\beta_K^*|\mathcal{M}^{'}\right)\\ &= \mathbb{E}\left(\frac{D^T X_K}{D^T D}\beta_K^*|\mathcal{M}^{'}\right)\end{aligned}$$

where $\mathcal{M}^{'} = E\cap\mathcal{E}_{t^*}$ such that $\mathbb{P}\left(\mathcal{M}^{'}\right) \geq 1-2\kappa-c_3^*k\exp(-c_4^*\log p)$ and the last line follows from the argument used to show (41).

The rest of argument is similar to what is used in showing 2. However, because (45) involves $D$, $\mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*|E,\,\mathcal{E}_{t^*}\right) \neq \mathbb{E}\left(\frac{\frac{1}{n}D^T X_K}{\frac{1}{n}D^T D}\beta_K^*|\hat{\gamma}_K = 0_k,\,\mathcal{E}_{t^*}\right)$, where $E =$

$\left\{\hat{\beta} = \hat{\gamma} = 0_p\right\}$. Instead of (33) (or, (42)) and (36), we apply

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{t_2}^{\prime}|E, \mathcal{E}_{t_1}\right) &\geq \mathbb{P}\left(\mathcal{E}_{t_2}^{\prime} \cap \mathcal{E}_{t_1} \cap E\right) \\
&\geq \mathbb{P}\left(\mathcal{E}_{t_1}\right) + \mathbb{P}\left(\mathcal{E}_{t_2}^{\prime}\right) + \mathbb{P}\left(E\right) - 2 \\
&\geq 1 - 2\kappa - c_1 \exp\left(-c_2 \log p\right) - k \exp\left(\frac{-nt_1^2}{2\frac{s}{n}\sigma_v^2}\right) \\
&\quad - \exp\left(\frac{-nt_2^2}{8\sigma_v^4}\right), \qquad \text{for any } t_2 \in (0,\, \sigma_v^2].
\end{aligned}$$

Consequently, we have the claim in Proposition 4.

Following the argument used to show Proposition 3, we also have the claim in Proposition 5.

### B.3 Simulations evidence

Here we evaluate the performance of the debiased Lasso proposed by van de Geer et al. (2014) based on the simulation setting of Section 3.2. Figures 13–14 present the results. Debiased Lasso exhibits substantial biases (relative to the standard deviation) and undercoverage for all values of $\sigma_x$ and its performance is very sensitive to the choice of $\lambda$. A comparison to the results in Section 3.2 shows that post double Lasso performs better than debiased Lasso.[12]

## C Lasso selection performance for different feasible penalty choices

In this section, investigate the selection performance of the Lasso for the three different penalty choices considered in Section 3.2: The heteroscedasticity-robust proposal

[12] We found that one of the reasons for the relatively poor performance of debiased Lasso is that $D$ is highly correlated with the important controls. Unreported simulation results show that debiased Lasso exhibits a better performance when $(D, X)$ exhibit a Toeplitz dependence structure as in the simulations reported by van de Geer et al. (2014).

Figure 13: Ratio of bias and standard deviation

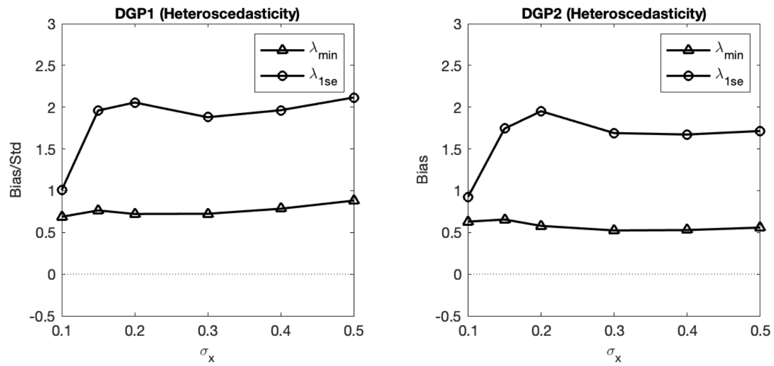
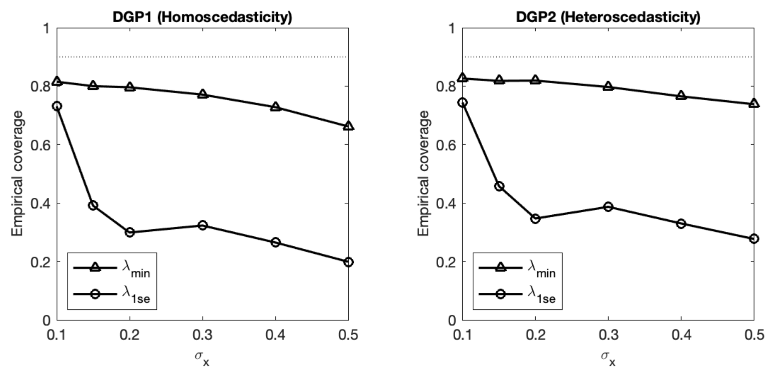



Figure 14: Coverage 90% confidence intervals




in Belloni et al. (2012) ($\lambda_{\text{BCCH}}$), the penalty parameter with the minimum cross-validated error ($\lambda_{\min}$), and the penalty parameter with the minimum cross-validation error plus one standard deviation ($\lambda_{1\text{se}}$).

We consider the following linear model:

$$Y_i = X_i\theta^* + \sigma(X_i)\varepsilon_i,$$

where $X_i \sim \mathcal{N}(0, \sigma_x^2 I_p)$ is independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ and $\{X_i, \varepsilon_i\}_{i=1}^n$ consists of i.i.d. entries. We set $n = 500$, $p = 200$, $\theta^* = (\underbrace{1, \dots, 1}_{k}, 0, \dots, 0)'$, and $k = 5$. We consider a homoscedastic DGP where $\sigma(X_i) = 1$ and a heteroscedastic DGP where $\sigma(X_i) = \sqrt{\frac{(1+X_i\gamma^*)^2}{\frac{1}{n}\sum_i (1+X_i\gamma^*)^2}}$. The results are based on 1,000 repetitions.

Figure 15 displays the average number of selected covariates as a function of $\sigma_x$. Lasso with $\lambda = \lambda_{\text{BCCH}}$ selects the lowest number of covariates. Choosing $\lambda = \lambda_{1\text{se}}$ leads to a somewhat higher number of selected covariates and results in moderate

Figure 15: Number of selected covariates

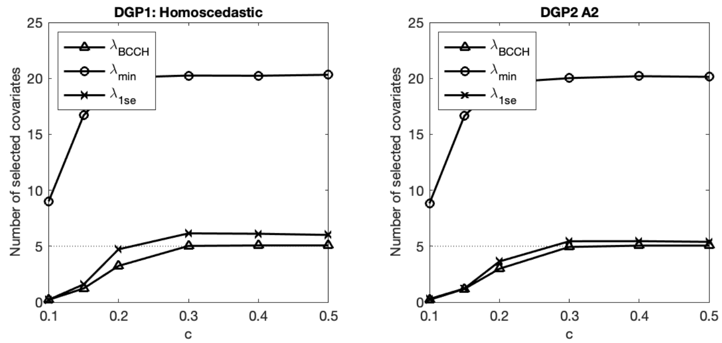
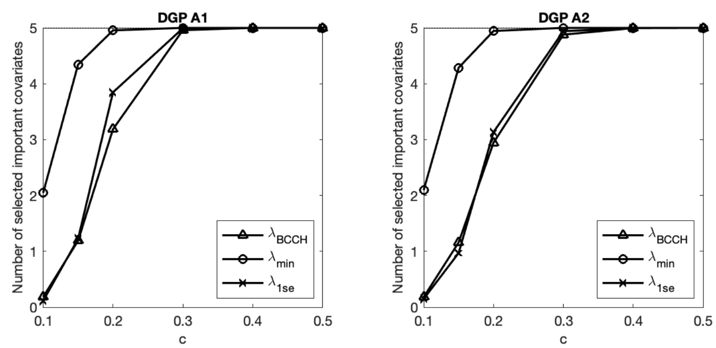



Figure 16: Number of selected relevant covariates




overselection for larger values of $\sigma_x$. Lasso with $\lambda = \lambda_{\text{min}}$ selects the highest number of covariates and exhibits substantial overselection. Figure 16 shows the corresponding numbers of selected relevant covariates. Lasso with $\lambda = \lambda_{\text{BCCH}}$ and $\lambda = \lambda_{\text{1se}}$ selects fewer relevant covariates than with $\lambda = \lambda_{\text{min}}$. We note that even when $\lambda = \lambda_{\text{min}}$, which result in substantial overselection, the Lasso is unable to select all the relevant regressors when $\sigma_x < 0.2$.

# D Additional simulations

In the main text, we consider a setting with normally distributed control variables, normally distributed errors terms, and $\alpha^* = 0$. Here we present additional simulation evidence based on the following DGP:

$$Y_i = D_i\alpha^* + X_i\beta^* + \eta_i, \tag{47}$$

$$D_i = X_i\gamma^* + v_i, \tag{48}$$

where $\eta_i$, $v_i$, and $X_i$ are independent of each other and $\{X_i, \eta_i, v_i\}_{i=1}^n$ consists of i.i.d. entries. The object of interest is $\alpha^*$. We set $n = 500$, $p = 200$, $\beta^* = \gamma^* = (\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)'$, and $k = 5$. We consider four DGPs that differ with respect to the distributions of $X_i$, $\eta_i$, and $v_i$, as well as $\alpha^*$. For DGP A1, we do not report the results for $\sigma_x < 0.2$ due to numerical issues with the computation of standard errors. The results are based on 1,000 simulation repetitions.

| | $X_i$ | $\eta_i$ | $v_i$ | $\alpha^*$ |
|---|---|---|---|---|
| DGP A1 | Indep. Bern $\left(\frac{1}{2}(1 - \sqrt{1 - 4\sigma_x^2})\right)$ | $\mathcal{N}(0,1)$ | $\mathcal{N}(0,1)$ | 0 |
| DGP A2 | $\mathcal{N}(0, \sigma_x^2 I_p)$ | $\frac{t(5)}{\sqrt{(5/3)}}$ | $\frac{t(5)}{\sqrt{(5/3)}}$ | 0 |
| DGP A3 | $\mathcal{N}(0, \sigma_x^2 I_p)$ | $\mathcal{N}(0,1)$ | $\mathcal{N}(0,1)$ | 1 |
| DGP A4 | $\mathcal{N}(0, \sigma_x^2 I_p)$ | $\mathcal{N}(0,1)$ | $\mathcal{N}(0,1)$ | -1 |

Figures 17–19 present the results. The most important determinant of the performance of post double Lasso is $\alpha^*$. To see why, recall that the reduced form parameter in the first step of post double Lasso (i.e., program (6)) is $\pi^* = \alpha^*\gamma^* + \beta^*$, which implies that the magnitude of $\pi^*$ depends on $\alpha^*$. Consequently, the selection performance of Lasso in the first step is directly affected by $\alpha^*$. In the extreme case where $\alpha^*$ is such that $\pi^* = 0_p$, Lasso does not select any controls if the penalty parameter is chosen according to the standard recommendations. The simulation results further show that there is no practical recommendation for choosing the penalty parameters. While $\lambda_{\min}$ leads to the best performance when $\alpha^* = 0$, this choice can yield poor performances when $\alpha^* \neq 0$. Finally, across all DGPs, OLS outperforms post double Lasso in terms of bias and coverage accuracy, but leads to somewhat wider confidence intervals.

Figure 17: Ratio of bias and standard deviation

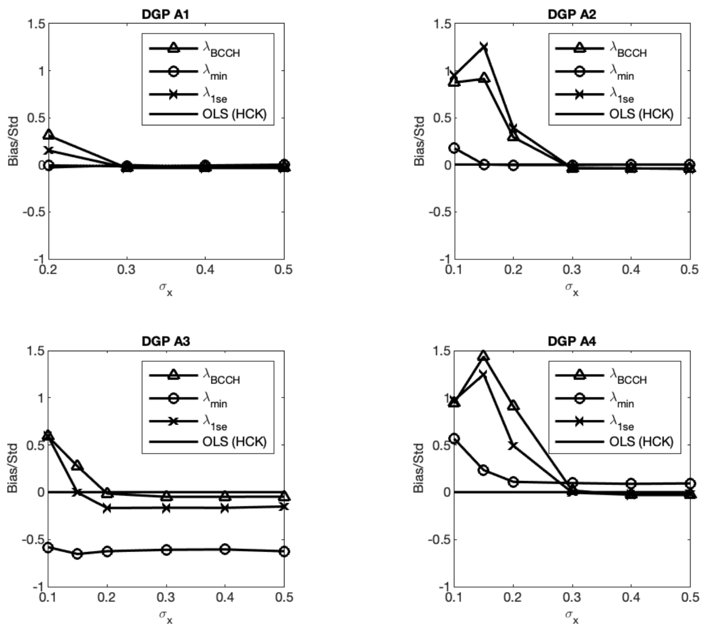



Figure 18: Coverage 90% confidence intervals

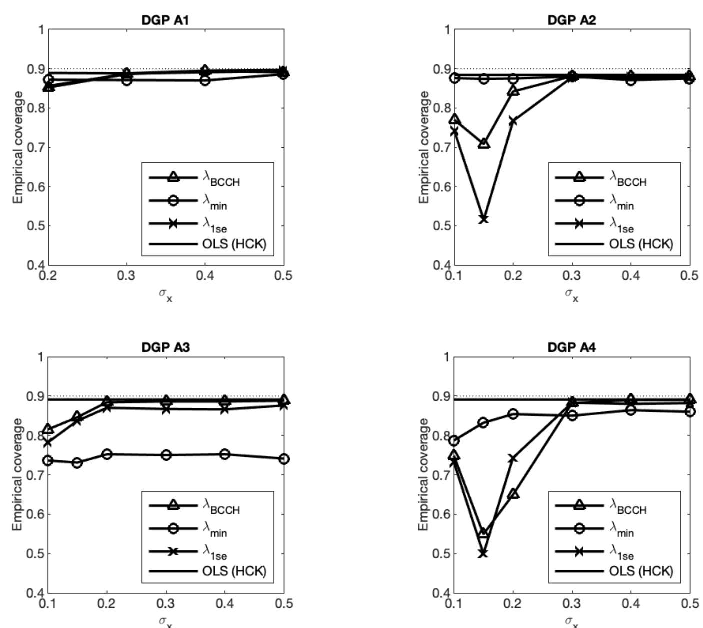
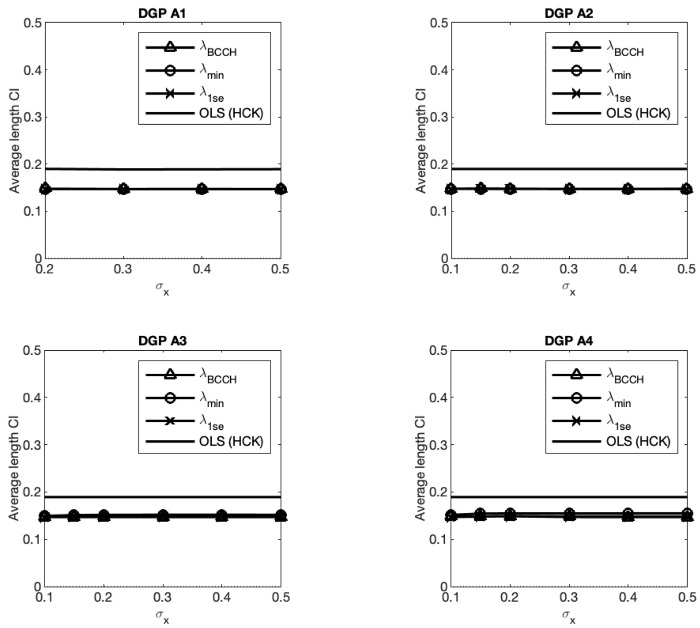


Figure 19: Average length 90% confidence intervals




# E Random design

## E.1 Results

In this section, we provide some results in Lemma 2 for the Lasso with a random design $X$. The necessary result on the Lasso's inclusion can be adopted in a similar fashion as in Propositions 2 and 4 to establish the OVBs.

We make the following assumption about (3).

**Assumption 3.** *Each row of $X$ is sampled independently; for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$, $\sup_{r\geq 1} r^{-\frac{1}{2}} \left(\mathbb{E}\left|X_{ij}\right|^r\right)^{\frac{1}{r}} \leq \alpha < \infty$; for any unit vector $a \in \mathbb{R}^k$ and $i = 1, \ldots, n$, $\sup_{r\geq 1} r^{-\frac{1}{2}} \left(\mathbb{E}\left|a^T X_{i,K}^T\right|^r\right)^{\frac{1}{r}} \leq \tilde{\alpha} < \infty$, where $X_{i,K}$ is the ith row of $X_K$ and $K = \left\{j : \theta_j^* \neq 0\right\}$. Moreover, the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent such that $\sup_{r\geq 1} r^{-\frac{1}{2}} \left(\mathbb{E}\left|\varepsilon_i\right|^r\right)^{\frac{1}{r}} \leq \sigma < \infty$ and $\mathbb{E}\left(X_i \varepsilon_i\right) = 0_p$ for all $i = 1, \ldots, n$.*

Assumption 3 is known as the sub-Gaussian tail condition defined in Vershynin (2012). Examples of sub-Gaussian variables include Gaussian mixtures and distributions with bounded support. The first and last part of Assumption 3 imply that

$X_{ij}, j = 1, \ldots, p$, and $\varepsilon_i$ are sub-Gaussian variables and is used in deriving the lower bounds on the regularization parameters. The second part of Assumption 3 is only used to establish some eigenvalue condition on $\frac{X_K^T X_K}{n}$.

**Assumption 4.** *The following conditions are satisfied: (i) $\theta^*$ is exactly sparse with at most $k$ non-zero coefficients and $K \neq \emptyset$; (ii)*

$$\left\| \left[ \mathbb{E}\left( X_{K^c}^T X_K \right) \right] \left[ \mathbb{E}\left( X_K^T X_K \right) \right]^{-1} \right\|_\infty = 1 - \phi \tag{49}$$

*for some $\phi \in (0,\, 1]$ such that $\phi^{-1} \precsim 1$; (iii) $\mathbb{E}(X_{ij}) = 0$ for all $j \in K$ and $\mathbb{E}\left(X_j^T X_j\right) \leq s$ for all $j = 1, \ldots, p$; (iv)*

$$\max\left\{ \frac{\phi}{12(1-\phi)k^{\frac{3}{2}}},\, \frac{\phi}{6k^{\frac{3}{2}}},\, \frac{\phi}{k} \right\} \sqrt{\frac{\log p}{n}} \;\leq\; \alpha^2 \quad \textit{if } \phi \in (0,\, 1)\,, \tag{50}$$

$$\max\left\{ \frac{1}{6k^{\frac{3}{2}}},\, \frac{1}{k} \right\} \sqrt{\frac{\log p}{n}} \;\leq\; \alpha^2 \quad \textit{if } \phi = 1, \tag{51}$$

$$\max\left\{ 2\tilde{\alpha}^2,\, 12\alpha^2,\, 1 \right\} \sqrt{\frac{\log p}{n}} \;\leq\; \lambda_{\min}\left( \mathbb{E}\left[ \frac{1}{n} X_K^T X_K \right] \right). \tag{52}$$

Part (iv) of Assumption 4 is imposed to ensure that

$$\left\| \left( \frac{1}{n} X_K^T X_K \right)^{-1} - \left[ \mathbb{E}\left( \frac{1}{n} X_K^T X_K \right) \right]^{-1} \right\|_\infty \;\precsim\; \frac{1}{\lambda_{\min}\left( \mathbb{E}\left[ \frac{1}{n} X_K^T X_K \right] \right)},$$

$$\left\| \frac{1}{n} X_{K^c}^T X_K \left( \frac{1}{n} X_K^T X_K \right)^{-1} \right\|_\infty \;\leq\; 1 - \frac{\phi}{2},$$

with high probability. To gain some intuition for (50)–(52), let us further assume $k \asymp 1$, $X_i$ is normally distributed for all $i = 1, \ldots, n$, and $\mathbb{E}\left(X_K^T X_K\right)$ is a diagonal matrix with the diagonal entries $\mathbb{E}\left(X_j^T X_j\right) = s \neq 0$. As a result, $\tilde{\alpha} = \alpha \asymp \sqrt{\frac{s}{n}}$ by the definition of a sub-Gaussian variable (e.g., Vershynin (2012)) and (50)–(52) essentially require $\sqrt{\frac{\log p}{n}} \precsim \frac{s}{n}$.

Given

$$\mathbb{P}\left( \left| \frac{X^T \varepsilon}{n} \right|_\infty \geq t \right) \leq 2 \exp\left( \frac{-nt^2}{c_0 \sigma^2 \alpha^2} + \log p \right). \tag{53}$$

and $\lambda \geq \frac{c\alpha\sigma\left(2 - \frac{\phi}{2}\right)}{\phi} \sqrt{\frac{\log p}{n}}$ for some sufficiently large universal constant $c > 0$, we have

$$\lambda \geq 2 \left| \frac{X^T \varepsilon}{n} \right|_\infty \tag{54}$$

with probability at least $1-c^{'}\exp\left(-c^{''}\log p\right)$.

Define the following events

$$
\begin{aligned}
\mathcal{E}_1 &= \left\{\left|\frac{X^T\varepsilon}{n}\right|_\infty \precsim \frac{\alpha\sigma\left(2-\frac{\phi}{2}\right)}{\phi}\sqrt{\frac{\log p}{n}}\right\}, \\
\mathcal{E}_2 &= \left\{\lambda_{\max}(\hat{\Sigma}_{KK}) \leq \frac{3}{2}\lambda_{\max}(\Sigma_{KK})\right\}, \\
\mathcal{E}_3 &= \left\{\left\|\left(\frac{1}{n}X_K^TX_K\right)^{-1} - \left[\mathbb{E}\left(\frac{1}{n}X_K^TX_K\right)\right]^{-1}\right\|_\infty \precsim \frac{1}{\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}X_K^TX_K\right]\right)}\right\}, \\
\mathcal{E}_4 &= \left\{\left\|\frac{1}{n}X_{K^c}^TX_K\left(\frac{1}{n}X_K^TX_K\right)^{-1}\right\|_\infty \leq 1-\frac{\phi}{2}\right\}.
\end{aligned}
$$

By (53), $\mathbb{P}\left(\mathcal{E}_1\right) \geq 1-c^{'}\exp\left(-c^{''}\log p\right)$; by (63), $\mathbb{P}\left(\mathcal{E}_2\right) \geq 1-c_1^{'}\exp\left(-c_1^{''}\log p\right)$; by (69), $\mathbb{P}\left(\mathcal{E}_3\right) \geq 1-c_2^{'}\exp\left(-c_2^{''}\left(\frac{\log p}{k^3}\right)\right)$; by (78), $\mathbb{P}\left(\mathcal{E}_4\right) \geq 1-c_3^{''}\exp\left(-b\left(\frac{\log p}{k^3}\right)\right)$, where $b$ is some positive constant that only depends on $\phi$ and $\alpha$.

**Lemma 2.** *Let Assumptions 3 and 4 hold. We solve the Lasso (4) with* $\lambda \geq \frac{c\alpha\sigma\left(2-\frac{\phi}{2}\right)}{\phi}\sqrt{\frac{\log p}{n}}$ *for some sufficiently large universal constant* $c>0$*. Suppose* $\mathbb{E}\left[X_K^TX_K\right]$ *is a positive definite matrix.*

*(i) Then, conditioning on* $\mathcal{E}_1\cap\mathcal{E}_4$ *(which holds with probability at least* $1-c_1\exp\left(-b\frac{\log p}{k^3}\right)$*), (4) has a unique optimal solution* $\hat{\theta}$ *such that* $\hat{\theta}_j = 0$ *for* $j \notin K$*.*

*(ii) With probability at least* $1-c_1\exp\left(-b\frac{\log p}{k^3}\right)$*,*

$$\left|\hat{\theta}_K-\theta_K^*\right|_2 \leq \frac{3\lambda\sqrt{k}}{\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}X_K^TX_K\right]\right)} \tag{55}$$

*where* $\theta_K = \left\{\theta_j\right\}_{j\in K}$ *and* $b$ *is some positive constant that only depends on* $\phi$ *and* $\alpha$*; if* $\mathbb{P}\left(\left\{supp(\hat{\theta})=K\right\}\cap\mathcal{E}_1\cap\mathcal{E}_2\right)>0$*, conditioning on* $\left\{supp(\hat{\theta})=K\right\}\cap\mathcal{E}_1\cap\mathcal{E}_2$*, we must have*

$$\left|\hat{\theta}_K-\theta_K^*\right|_2 \geq \frac{\lambda\sqrt{k}}{3\lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n}X_K^TX_K\right]\right)} \geq \frac{\lambda\sqrt{k}}{3\sum_{j\in K}\left(\mathbb{E}\left[\frac{1}{n}X_j^TX_j\right]\right)}. \tag{56}$$

*(iii) If* $\mathbb{E}\left(X_K^TX_K\right)$ *is a diagonal matrix with the diagonal entries* $\mathbb{E}\left(X_j^TX_j\right)=s\neq 0$*, then*

$$\left|\hat{\theta}_j-\theta_j^*\right| \leq \frac{7\lambda n}{4s} \qquad \forall j\in K \tag{57}$$

*with probability at least* $1-c_1\exp\left(-b\frac{\log p}{k^3}\right)$*; if* $\mathbb{P}\left(\left\{\hat{\theta}_j\neq 0,\, j\in K\right\}\cap\mathcal{E}_1\cap\mathcal{E}_3\cap\mathcal{E}_4\right)>0$*, conditioning on* $\left\{\hat{\theta}_j\neq 0,\, j\in K\right\}\cap\mathcal{E}_1\cap\mathcal{E}_3\cap\mathcal{E}_4$*, we must have*

$$\left|\hat{\theta}_j-\theta_j^*\right|\geq\frac{\lambda n}{4s}\geq c_0\frac{\sigma}{\phi}\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}. \tag{58}$$

*(iv) Suppose* $K=\{1\}$ *and* $\mathbb{E}\left(X_1^TX_1\right)=s\neq 0$*. If*

$$|\theta_1^*|\leq\frac{\lambda n}{4s}, \tag{59}$$

*then we must have*

$$\mathbb{P}\left(\hat{\theta}=0_p\right)\geq 1-c\exp\left(-b\log p\right). \tag{60}$$

The part $\frac{\lambda n}{4s}\geq c_0\frac{\sigma}{\phi}\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$ in bound (58) follows from the fact that $\alpha\gtrsim\sqrt{\frac{s}{n}}=\sqrt{\mathbb{E}\left(\frac{1}{n}\sum X_{ij}^2\right)}$ where $j\in K$. If we further assume $k\asymp 1$ and $X_i$ is normally distributed for all $i=1,\ldots,n$, then $\tilde{\alpha}=\alpha\asymp\sqrt{\frac{s}{n}}$ and (50)-(52) imply that $\sqrt{\frac{\log p}{n}}\precsim\frac{s}{n}$ (recalling our discussion following Assumption 4). Taking the worst case $\sqrt{\frac{\log p}{n}}\asymp\frac{s}{n}$ and the optimal choice $\lambda\asymp\sqrt{\frac{s}{n}}\sqrt{\frac{\log p}{n}}$, (58) implies that

$$\left|\hat{\theta}_j-\theta_j^*\right|\gtrsim\sigma\left(\frac{\log p}{n}\right)^{\frac{1}{4}}$$

conditioning on $\left\{\hat{\theta}_j\neq 0,\, j\in K\right\}\cap\mathcal{E}_1\cap\mathcal{E}_3\cap\mathcal{E}_4$; therefore, the minimax rate $\sqrt{\frac{k\log\frac{p}{k}}{n}}\asymp\sqrt{\frac{\log p}{n}}$ (given $k\asymp 1$) can no longer be attained by the Lasso. In the example where $K=1$, if (59) is satisfied, then Lasso sets $\hat{\theta}_1=0$ with high probability.

## E.2 Main proof for Lemma 2

In what follows, we let $\Sigma_{KK}:=\mathbb{E}\left[\frac{1}{n}X_K^TX_K\right]$, $\hat{\Sigma}_{KK}:=\frac{1}{n}X_K^TX_K$, and $\lambda_{\min}\left(\Sigma\right)$ denote the minimum eigenvalue of the matrix $\Sigma$. The proof for Proposition 2(i) follows similar argument as before but requires a few extra steps. In applying Lemma 7.23 from Chapter 7.5 of Wainwright (2019) to establish the uniqueness of $\hat{\theta}$ upon the success of PDW construction, it suffices to show that $\lambda_{\min}(\hat{\Sigma}_{KK})\geq\frac{1}{2}\lambda_{\min}(\Sigma_{KK})$ and this fact is verified in (72) in the appendix. As a consequence, the subproblem (25) is strictly convex and has a unique minimizer. The details that show the PDW construction succeeds conditioning on $\mathcal{E}_1 \cap \mathcal{E}_4$ (which holds with probability at least $1 - c_1 \exp\left(-b\frac{\log p}{k^3}\right)$) can be found in Lemma 6 (where $b$ is some positive constant that only depends on $\phi$ and $\alpha$).

To show (55), note that our choice of $\lambda$ and $\left|\hat{\delta}_K\right| \leq 1$ yield

$$|\Delta| \leq \left|\lambda\hat{\delta}_K\right| + \left|\frac{X_K^T\varepsilon}{n}\right| \leq \frac{3\lambda}{2}1_k,$$

which implies that $|\Delta|_2 \leq \frac{3\lambda}{2}\sqrt{k}$. Moreover, we can show

$$\begin{aligned} \left|\hat{\theta}_K - \theta_K^*\right|_2 &= \frac{\left|\left(\frac{X_K^TX_K}{n}\right)^{-1}\Delta\right|_2}{|\Delta|_2}|\Delta|_2 \\ &\leq \frac{1}{\lambda_{\min}\left(\frac{1}{n}X_K^TX_K\right)}\frac{3\lambda}{2}\sqrt{k}. \end{aligned} \tag{61}$$

Applying (72) and the bound $|\Delta|_2 \leq \frac{3\lambda}{2}\sqrt{k}$ yields the claim.

In showing (56) in (ii) and (58) in (iii), we will condition on $\left\{\text{supp}(\hat{\theta}) = K\right\} \cap \mathcal{E}_1 \cap \mathcal{E}_2$ and $\left\{\hat{\theta}_j \neq 0,\, j \in K\right\} \cap \mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, respectively.

To show (56), note that in Step 2 of the PDW procedure, $\hat{\delta}_K$ is chosen such that $\left|\hat{\delta}_j\right| = 1$ for any $j \in K$ whenever $\text{supp}(\hat{\theta}) = K$. Given the choice of $\lambda$, we are ensured to have

$$|\Delta| \geq \left|\left|\lambda\hat{\delta}_K\right| - \left|\frac{X_K^T\varepsilon}{n}\right|\right| \geq \frac{\lambda}{2}1_k,$$

which implies that $|\Delta|_2 \geq \frac{\lambda}{2}\sqrt{k}$. Moreover, we can show

$$\left|\hat{\theta}_K - \theta_K^*\right|_2 = \frac{\left|\left(\frac{X_K^TX_K}{n}\right)^{-1}\Delta\right|_2}{|\Delta|_2}|\Delta|_2 \geq \frac{1}{\lambda_{\max}\left(\frac{1}{n}X_K^TX_K\right)}\frac{\lambda}{2}\sqrt{k}. \tag{62}$$

It remains to bound $\lambda_{\max}\left(\hat{\Sigma}_{KK}\right)$. We first write

$$\begin{aligned} \lambda_{\max}(\Sigma_{KK}) &= \max_{||h'||_2=1} \mu'^T\Sigma_{KK}\mu' \\ &= \max_{||h'||_2=1} \left[\mu'^T\hat{\Sigma}_{KK}\mu' + \mu'^T(\Sigma_{KK} - \hat{\Sigma}_{KK})\mu'\right] \\ &\geq \mu^T\hat{\Sigma}_{KK}\mu + \mu^T(\Sigma_{KK} - \hat{\Sigma}_{KK})\mu \end{aligned}$$

where $\mu \in \mathbb{R}^k$ is a unit-norm maximal eigenvector of $\hat{\Sigma}_{KK}$. Applying Lemma 3(b) with $t = \tilde{\alpha}^2 \sqrt{\frac{\log p}{n}}$ yields

$$\mu^T \left( \Sigma_{KK} - \hat{\Sigma}_{KK} \right) \mu \geq -\tilde{\alpha}^2 \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, provided that $\sqrt{\frac{\log p}{n}} \leq 1$; therefore, $\lambda_{\max}(\Sigma_{KK}) \geq \lambda_{\max}(\hat{\Sigma}_{KK}) - \tilde{\alpha}^2 \sqrt{\frac{\log p}{n}}$. Because $\tilde{\alpha}^2 \sqrt{\frac{\log p}{n}} \leq \frac{\lambda_{\max}(\Sigma_{KK})}{2}$ (implied by (52)), we have

$$\lambda_{\max}(\hat{\Sigma}_{KK}) \leq \frac{3}{2} \lambda_{\max}(\Sigma_{KK}) \tag{63}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

As a consequence,

$$\left| \hat{\theta}_K - \theta_K^* \right|_2 \geq \frac{1}{\lambda_{\max}\left( \frac{1}{n} X_K^T X_K \right)} \frac{\lambda}{2} \sqrt{k} \geq \frac{1}{\lambda_{\max}(\Sigma_{KK})} \frac{\lambda}{3} \sqrt{k}.$$

The second inequality in (56) simply follows from the fact $\lambda_{\max}\left( \mathbb{E}\left[ \frac{1}{n} X_K^T X_K \right] \right) \leq \sum_{j \in K} \left( \mathbb{E}\left[ \frac{1}{n} X_j^T X_j \right] \right)$.

To show (57), note that

$$\begin{aligned} \left| \hat{\theta}_K - \theta_K^* \right|_\infty &\leq \left| \hat{\Sigma}_{KK}^{-1} \frac{X_K^T \varepsilon}{n} \right|_\infty + \lambda \left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty \\ &\leq \left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty \left| \frac{X_K^T \varepsilon}{n} \right|_\infty + \lambda \left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty \\ &\leq \frac{3\lambda}{2} \left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty . \end{aligned} \tag{64}$$

We then apply (69) of Lemma 4 in the appendix, and the fact $\left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty - \left\| \Sigma_{KK}^{-1} \right\|_\infty \leq \left\| \hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1} \right\|_\infty$ (so that $\left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty \leq \frac{7n}{6s}$); putting everything yields the claim.

To show (58), we again carry over the argument in the proof for Lemma 1. Letting $M = \hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}$, we have

$$\begin{aligned} \left| \hat{\theta}_K - \theta_K^* \right| &= \left| \left( \Sigma_{KK}^{-1} + M \right) \left[ \left( \frac{X_K^T \varepsilon}{n} \right) - \lambda \hat{\delta}_K \right] \right| \\ &\geq \left| \Sigma_{KK}^{-1} \left[ \left( \frac{X_K^T \varepsilon}{n} \right) + \lambda \hat{\delta}_K \right] \right| - \left| M \left[ \left( \frac{X_K^T \varepsilon}{n} \right) - \lambda \hat{\delta}_K \right] \right| \\ &\geq \left| \Sigma_{KK}^{-1} \right| \left| \left| \lambda \hat{\delta}_K \right| - \left| \frac{X_K^T \varepsilon}{n} \right| \right| - \| M \|_\infty \left| \left( \frac{X_K^T \varepsilon}{n} \right) - \lambda \hat{\delta}_K \right|_\infty 1_k, \end{aligned}$$

where the third line uses the fact that $\Sigma_{KK}^{-1}$ is diagonal.

Note that as before, the choice of $\lambda$ stated in Lemma 2 and the fact $\Sigma_{KK}^{-1} = \frac{n}{s} I_k$ yield

$$\begin{aligned}\left|\hat{\theta}_j - \theta_j^*\right| &\geq \frac{\lambda n}{2s} - \|M\|_\infty \left|\left(\frac{X_K^T \varepsilon}{n}\right) - \lambda \hat{\delta}_K\right|_\infty \\ &\geq \frac{\lambda n}{2s} - \frac{3}{2}\lambda \|M\|_\infty .\end{aligned}$$

By (69) of Lemma 4 in the appendix, with probability at least $1 - c_1 \exp\left(-b\frac{\log p}{k^3}\right)$, $\|M\|_\infty \leq \frac{1}{6}\lambda_{\min}^{-1}(\Sigma_{KK}) = \frac{n}{6s}$.

As a result, we have (58). The part $\frac{\lambda n}{4s} \geq c_0 \frac{\sigma}{\phi}\sqrt{\frac{n}{s}}\sqrt{\frac{\log p}{n}}$ in bound (58) follows from the fact that $\alpha \gtrsim \sqrt{\frac{s}{n}} = \sqrt{\mathbb{E}\left(\frac{1}{n}\sum X_{ij}^2\right)}$ where $j \in K$.

To establish (60), we adopt argument similar to what is used in showing (18) by applying the KKT condition

$$\left(\frac{1}{n} X_1^T X_1\right)\left(\theta_1^* - \hat{\theta}_1\right) = \lambda \mathrm{sgn}\left(\hat{\theta}_1\right) - \frac{X_1^T \varepsilon}{n}$$

and defining $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_4$.

## E.3 Additional technical lemmas and proofs

In this section, we show that the PDW construction succeeds with high probability in Lemma 6, which is proved using results from Lemmas 3–5. The derivations for Lemmas 4 and 5 modify the argument in Wainwright (2009) and Ravikumar et al. (2010) to make it suitable for our purposes. In what follows, we let $\Sigma_{K^cK} := \mathbb{E}\left[\frac{1}{n} X_{K^c}^T X_K\right]$ and $\hat{\Sigma}_{K^cK} := \frac{1}{n} X_{K^c}^T X_K$. Similarly, let $\Sigma_{KK} := \mathbb{E}\left[\frac{1}{n} X_K^T X_K\right]$ and $\hat{\Sigma}_{KK} := \frac{1}{n} X_K^T X_K$.

**Lemma 3.** *(a) Let $(W_i)_{i=1}^n$ and $\left(W_i^{'}\right)_{i=1}^n$ consist of independent components, respectively. Suppose there exist parameters $\alpha$ and $\alpha^{'}$ such that*

$$\begin{aligned}\sup_{r\geq 1} r^{-\frac{1}{2}} \left(\mathbb{E}\left|W_i\right|^r\right)^{\frac{1}{r}} &\leq \alpha, \\ \sup_{r\geq 1} r^{-\frac{1}{2}} \left(\mathbb{E}\left|W_i^{'}\right|^r\right)^{\frac{1}{r}} &\leq \alpha^{'},\end{aligned}$$

*for all* $i = 1, \ldots, n$. *Then*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(W_i W_i^{'}\right) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(W_i W_i^{'}\right)\right]\right| \geq t\right] \leq 2\exp\left(-cn\left(\frac{t^2}{\alpha^2\alpha'^2} \wedge \frac{t}{\alpha\alpha^{'}}\right)\right). \tag{65}$$

*(b) For any unit vector* $v \in \mathbb{R}^d$, *suppose there exists a parameter* $\tilde{\alpha}$ *such that*

$$\sup_{r\geq 1} r^{-\frac{1}{2}}\left(\mathbb{E}\left|a^T Z_i^T\right|^r\right)^{\frac{1}{r}} \leq \tilde{\alpha},$$

*where* $Z_i$ *is the ith row of* $Z \in \mathbb{R}^{n\times d}$, *then we have*

$$\mathbb{P}(\left||Zv|_2^2 - \mathbb{E}\left(|Zv|_2^2\right)\right| \geq nt) \leq 2\exp\left(-c^{'}n\left(\frac{t^2}{\tilde{\alpha}^4} \wedge \frac{t}{\tilde{\alpha}^2}\right)\right).$$

**Remark 6.** *Lemma 3 is based on Lemma 5.14 and Corollary 5.17 in Vershynin (2012).*

**Lemma 4.** *Suppose Assumption 3 holds. For any* $t > 0$ *and some constant* $c > 0$, *we have*

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty \geq t\right\} \leq 2(p-k)k\exp\left(-cn\left(\frac{t^2}{k^2\alpha^4} \wedge \frac{t}{k\alpha^2}\right)\right), \tag{66}$$

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \geq t\right\} \leq 2k^2\exp\left(-cn\left(\frac{t^2}{k^2\alpha^4} \wedge \frac{t}{k\alpha^2}\right)\right). \tag{67}$$

*Furthermore, if* $k \geq 1$, $\frac{\log p}{n} \leq 1$, $\tilde{\alpha}^2\sqrt{\frac{\log p}{n}} \leq \frac{\lambda_{\min}(\Sigma_{KK})}{2}$, *and* $\alpha^2\sqrt{\frac{\log p}{n}} \leq \frac{\lambda_{\min}(\Sigma_{KK})}{12}$, *we have*

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma_{KK})}\right\} \geq 1 - c_1^{'}\exp\left(-c_2^{'}\log p\right), \tag{68}$$

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty \leq \frac{1}{6\lambda_{\min}(\Sigma_{KK})}\right\} \geq 1 - c_1\exp\left(-c_2\left(\frac{\log p}{k^3}\right)\right). \tag{69}$$

*Proof.* Let $u_{j'j}$ denote the element $(j^{'}, j)$ of the matrix difference $\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}$. The definition of the $l_\infty$matrix norm implies that

$$\begin{aligned}\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty \geq t\right\} &= \mathbb{P}\left\{\max_{j'\in K^c}\sum_{j\in K}|u_{j'j}|\geq t\right\}\\ &\leq (p-k)\mathbb{P}\left\{\sum_{j\in K}|u_{j'j}|\geq t\right\}\\ &\leq (p-k)\mathbb{P}\left\{\exists j\in K\,|\,|u_{j'j}|\geq \frac{t}{k}\right\}\\ &\leq (p-k)k\mathbb{P}\left\{|u_{j'j}|\geq \frac{t}{k}\right\}\\ &\leq (p-k)k\cdot 2\exp\left(-cn\left(\frac{t^2}{k^2\alpha^4}\wedge\frac{t}{k\alpha^2}\right)\right),\end{aligned}$$

where the last inequality follows Lemma 3(a). Bound (67) can be derived in a similar fashion except that the pre-factor $(p-k)$ is replaced by $k$.

To prove (69), note that

$$\begin{aligned}\left\|\hat{\Sigma}_{KK}^{-1}-\Sigma_{KK}^{-1}\right\|_\infty &= \left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK}-\hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_\infty\\ &\leq \sqrt{k}\left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK}-\hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_2\\ &\leq \sqrt{k}\left\|\Sigma_{KK}^{-1}\right\|_2\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2\\ &\leq \frac{\sqrt{k}}{\lambda_{\min}(\Sigma_{KK})}\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2. \qquad (70)\end{aligned}$$

To bound $\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2$ in (70), we apply (67) with $t=\frac{\alpha^2}{\sqrt{k}}\sqrt{\frac{\log p}{n}}$ and obtain

$$\left\|\hat{\Sigma}_{KK}-\Sigma_{KK}\right\|_2\leq\frac{\alpha^2}{\sqrt{k}}\sqrt{\frac{\log p}{n}},$$

with probability at least $1-c_1\exp\left(-c_2\frac{\log p}{k^3}\right)$, provided that $k^{-3}\frac{\log p}{n}\leq 1$. To bound $\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2$ in (70), let us write

$$\begin{aligned}\lambda_{\min}(\Sigma_{KK}) &= \min_{||\mu'||_2=1}\mu'^T\Sigma_{KK}\mu'\\ &= \min_{||\mu'||_2=1}\left[\mu'^T\hat{\Sigma}_{KK}\mu'+\mu'^T(\Sigma_{KK}-\hat{\Sigma}_{KK})\mu'\right]\\ &\leq \mu^T\hat{\Sigma}_{KK}\mu+\mu^T(\Sigma_{KK}-\hat{\Sigma}_{KK})\mu \qquad (71)\end{aligned}$$

where $\mu \in \mathbb{R}^k$ is a unit-norm minimal eigenvector of $\hat{\Sigma}_{KK}$. We then apply Lemma 3(b) with $t = \tilde{\alpha}^2\sqrt{\frac{\log p}{n}}$ to show

$$\left|\mu^T\left(\Sigma_{KK} - \hat{\Sigma}_{KK}\right)\mu\right| \le \tilde{\alpha}^2\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1^{'}\exp\left(-c_2^{'}\log p\right)$, provided that $\sqrt{\frac{\log p}{n}} \le 1$. Therefore, $\lambda_{\min}(\Sigma_{KK}) \le \lambda_{\min}(\hat{\Sigma}_{KK}) + \tilde{\alpha}^2\sqrt{\frac{\log p}{n}}$. As long as $\tilde{\alpha}^2\sqrt{\frac{\log p}{n}} \le \frac{\lambda_{\min}(\Sigma_{KK})}{2}$, we have

$$\lambda_{\min}(\hat{\Sigma}_{KK}) \ge \frac{1}{2}\lambda_{\min}(\Sigma_{KK}), \tag{72}$$

and consequently (68),

$$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \le \frac{2}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1 - c_1^{'}\exp\left(-c_2^{'}\log p\right)$.

Putting the pieces together, as long as $\frac{\alpha^2}{\lambda_{\min}(\Sigma_{KK})}\sqrt{\frac{\log p}{n}} \le \frac{1}{12}$,

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty \le \frac{\sqrt{k}}{\lambda_{\min}(\Sigma_{KK})}\frac{\alpha^2}{\sqrt{k}}\sqrt{\frac{\log p}{n}}\frac{2}{\lambda_{\min}(\Sigma_{KK})} \le \frac{1}{6\lambda_{\min}(\Sigma_{KK})} \tag{73}$$

with probability at least $1 - c_1\exp\left(-c_2\frac{\log p}{k^3}\right)$. □

**Lemma 5.** *Let Assumption 3 hold. Suppose*

$$\left\|\mathbb{E}\left[X_{K^c}^T X_K\right]\left[\mathbb{E}(X_K^T X_K)\right]^{-1}\right\|_\infty = 1 - \phi \tag{74}$$

*for some $\phi \in (0,\,1]$. If $k \ge 1$ and*

$$\max\left\{\frac{\phi}{12(1-\phi)k^{\frac{3}{2}}},\, \frac{\phi}{6k^{\frac{3}{2}}},\, \frac{\phi}{k}\right\}\sqrt{\frac{\log p}{n}} \;\le\; \alpha^2 \quad \textit{if } \phi \in (0,\,1)\,, \tag{75}$$

$$\max\left\{\frac{1}{6k^{\frac{3}{2}}},\, \frac{1}{k}\right\}\sqrt{\frac{\log p}{n}} \;\le\; \alpha^2 \quad \textit{if } \phi = 1, \tag{76}$$

$$\max\left\{2\tilde{\alpha}^2,\, 12\alpha^2,\, 1\right\}\sqrt{\frac{\log p}{n}} \;\le\; \lambda_{\min}(\Sigma_{KK}), \tag{77}$$

*then for some positive constant $b$ that only depends on $\phi$ and $\alpha$, we have*

$$\mathbb{P}\left[\left\|\frac{1}{n}X_{K^c}^T X_K\left(\frac{1}{n}X_K^T X_K\right)^{-1}\right\|_\infty \ge 1 - \frac{\phi}{2}\right] \le c^{'}\exp\left(-b\left(\frac{\log p}{k^3}\right)\right). \tag{78}$$

*Proof.* Using the decomposition in Ravikumar et al. (2010), we have

$$\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1} - \Sigma_{K^cK}\Sigma_{KK}^{-1} = R_1 + R_2 + R_3,$$

where

$$\begin{aligned} R_1 &= \Sigma_{K^cK}\left[\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right], \\ R_2 &= \left[\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right]\Sigma_{KK}^{-1}, \\ R_3 &= \left[\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right]\left[\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right]. \end{aligned}$$

By (74), we have $\left\|\Sigma_{K^cK}\Sigma_{KK}^{-1}\right\|_\infty = 1-\phi$. It suffices to show $\|R_i\|_\infty \le \frac{\phi}{6}$ for $i = 1, ..., 3$.

For $R_1$, note that

$$R_1 = -\Sigma_{K^cK}\Sigma_{KK}^{-1}[\hat{\Sigma}_{KK} - \Sigma_{KK}]\hat{\Sigma}_{KK}^{-1}.$$

Applying the facts $\|AB\|_\infty \le \|A\|_\infty \|B\|_\infty$ and $\|A\|_\infty \le \sqrt{a}\,\|A\|_2$ for any symmetric matrix $A \in \mathbb{R}^{a\times a}$, we can bound $R_1$ in the following fashion:

$$\begin{aligned} \|R_1\|_\infty &\le \left\|\Sigma_{K^cK}\Sigma_{KK}^{-1}\right\|_\infty \left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \left\|\hat{\Sigma}_{KK}^{-1}\right\|_\infty \\ &\le (1-\phi)\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \sqrt{k}\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2, \end{aligned}$$

where the last inequality uses (74). If $\phi = 1$, then $\|R_1\|_\infty = 0$ so we may assume $\phi < 1$ in the following. Bound (68) from the proof for Lemma 4 yields

$$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \le \frac{2}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1-c_1\exp(-c_2\log p)$. Now, we apply bound (67) from Lemma 4 with $t = \frac{\phi}{12(1-\phi)}\sqrt{\frac{\log p}{kn}}$ and obtain

$$\mathbb{P}\left[\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \ge \frac{\phi}{12(1-\phi)}\sqrt{\frac{\log p}{kn}}\right] \le 2\exp\left(-c\left(\frac{\phi^2\log p}{\alpha^4(1-\phi)^2k^3}\right)\right),$$

provided $\frac{\phi}{12(1-\phi)\alpha^2 k}\sqrt{\frac{\log p}{kn}} \le 1$. Then, if $\sqrt{\frac{\log p}{n}} \le \lambda_{\min}(\Sigma_{KK})$, we are guaranteed that

$$\mathbb{P}\left[\|R_1\|_\infty \ge \frac{\phi}{6}\right] \le 2\exp\left(-c\left(\frac{\phi^2\log p}{\alpha^4(1-\phi)^2k^3}\right)\right) + c_1\exp(-c_2\log p).$$

For $R_2$, note that

$$\begin{aligned} \|R_2\|_\infty &\leq \sqrt{k}\left\|\Sigma_{KK}^{-1}\right\|_2\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty \\ &\leq \frac{\sqrt{k}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty . \end{aligned}$$

If $\frac{\phi}{6\alpha^2 k}\sqrt{\frac{\log p}{kn}} \leq 1$ and $\sqrt{\frac{\log p}{n}} \leq \lambda_{\min}(\Sigma_{KK})$, applying bound (66) from Lemma 4 with $t = \frac{\phi}{6}\sqrt{\frac{\log p}{kn}}$ yields

$$\mathbb{P}\left[\|R_2\|_\infty \geq \frac{\phi}{6}\right] \leq 2\exp\left(-c\left(\frac{\phi^2\log p}{\alpha^4 k^3}\right)\right).$$

For $R_3$, applying (66) with $t = \phi\sqrt{\frac{\log p}{n}}$ to bound $\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty$ and (69) to bound $\left\|\hat{\Sigma}_{KK}^{-1}-\Sigma_{KK}^{-1}\right\|_\infty$ yields

$$\mathbb{P}\left[\|R_3\|_\infty \geq \frac{\phi}{6}\right] \leq c^{'}\left[\exp\left(-c\left(\frac{\phi^2\log p}{\alpha^4 k^3}\right)\right)+\exp\left(-c\left(\frac{\log p}{k^3}\right)\right)\right],$$

provided that $\frac{\phi}{\alpha^2 k}\sqrt{\frac{\log p}{n}} \leq 1$ and $\sqrt{\frac{\log p}{n}} \leq \lambda_{\min}(\Sigma_{KK})$.

Putting everything together, we conclude that

$$\mathbb{P}\left[\left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty \geq 1-\frac{\phi}{2}\right] \leq c^{'}\exp\left(-b\left(\frac{\log p}{k^3}\right)\right)$$

for some positive constant $b$ that only depends on $\phi$ and $\alpha$. □

**Lemma 6.** *Let the assumptions in Lemmas 4 and 5 hold. Suppose $\theta^*$ is exactly sparse with at most $k$ non-zero coefficients and $K = \left\{j : \theta_j^* \neq 0\right\} \neq \emptyset$. If we choose $\lambda \geq \frac{c\alpha\sigma\left(2-\frac{\phi}{2}\right)}{\phi}\sqrt{\frac{\log p}{n}}$ for some sufficiently large universal constant $c > 0$, $\left|\hat{\delta}_{K^c}\right|_\infty \leq 1-\frac{\phi}{4}$ with probability at least $1-c_1\exp\left(-b\frac{\log p}{k^3}\right)$, where $b$ is some positive constant that only depends on $\phi$ and $\alpha$.*

*Proof.* By construction, the subvectors $\hat{\theta}_K$, $\hat{\delta}_K$, and $\hat{\delta}_{K^c}$ satisfy the zero-subgradient condition in the PDW construction. With the fact that $\hat{\theta}_{K^c} = \theta_{K^c}^* = 0_{p-k}$, we have

$$\begin{aligned} \hat{\Sigma}_{KK}\left(\hat{\theta}_K-\theta_K^*\right)-\frac{1}{n}X_K^T\varepsilon+\lambda\hat{\delta}_K &= 0_k, \\ \hat{\Sigma}_{K^cK}\left(\hat{\theta}_K-\theta_K^*\right)-\frac{1}{n}X_{K^c}^T\varepsilon+\lambda\hat{\delta}_{K^c} &= 0_{p-k}. \end{aligned}$$

The equations above yields

$$\begin{aligned}\hat{\delta}_{K^c} &= -\frac{1}{\lambda}\hat{\Sigma}_{K^cK}\left(\hat{\theta}_K - \theta_K^*\right) + X_{K^c}^T\frac{\varepsilon}{n\lambda},\\ \hat{\theta}_K - \theta_K^* &= \hat{\Sigma}_{KK}^{-1}\frac{X_K^T\varepsilon}{n} - \lambda\hat{\Sigma}_{KK}^{-1}\hat{\delta}_K,\end{aligned}$$

which yields

$$\hat{\delta}_{K^c} = \left(\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right)\hat{\delta}_K + \left(X_{K^c}^T\frac{\varepsilon}{n\lambda}\right) - \left(\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right)X_K^T\frac{\varepsilon}{n\lambda}.$$

Using elementary inequalities and the fact that $\left|\hat{\delta}_K\right|_\infty \leq 1$, we obtain

$$\left|\hat{\delta}_{K^c}\right|_\infty \leq \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty + \left|X_{K^c}^T\frac{\varepsilon}{n\lambda}\right|_\infty + \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty\left|X_K^T\frac{\varepsilon}{n\lambda}\right|_\infty.$$

By Lemma 5, $\left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty \leq 1 - \frac{\phi}{2}$ with probability at least $1 - c^{'}\exp\left(-\frac{b\log p}{k^3}\right)$; as a result,

$$\begin{aligned}\left|\hat{\delta}_{K^c}\right|_\infty &\leq 1 - \frac{\phi}{2} + \left|X_{K^c}^T\frac{\varepsilon}{n\lambda}\right|_\infty + \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty\left|X_K^T\frac{\varepsilon}{n\lambda}\right|_\infty\\ &\leq 1 - \frac{\phi}{2} + \left(2 - \frac{\phi}{2}\right)\left|\hat{X}^T\frac{\varepsilon}{n\lambda}\right|_\infty.\end{aligned}$$

It remains to show that $\left(2 - \frac{\phi}{2}\right)\left|X^T\frac{\varepsilon}{n\lambda}\right|_\infty \leq \frac{\phi}{4}$ with high probability. This result holds if $\lambda \geq \frac{4\left(2-\frac{\phi}{2}\right)}{\phi}\left|X^T\frac{\varepsilon}{n}\right|_\infty$. In particular, Lemma 3(a) and a union bound imply that

$$\mathbb{P}\left(\left|\frac{X^T\varepsilon}{n}\right|_\infty \geq t\right) \leq 2\exp\left(\frac{-nt^2}{c_0\sigma^2\alpha^2} + \log p\right).$$

Thus, under the choice of $\lambda$ in Lemma 6, we have $\left|\hat{\delta}_{K^c}\right|_\infty \leq 1 - \frac{\phi}{4}$ with probability at least $1 - c_1\exp\left(-b\frac{\log p}{k^3}\right)$. □