OMITTED VARIABLE BIAS OF LASSO-BASED INFERENCE METHODS: A FINITE SAMPLE ANALYSIS

Kaspar Wüthrich and Ying Zhu*

Abstract—We study the finite sample behavior of Lasso-based inference methods such as post–double Lasso and debiased Lasso. We show that these methods can exhibit substantial omitted variable biases (OVBs) due to Lasso’s not selecting relevant controls. This phenomenon can occur even when the coefficients are sparse and the sample size is large and larger than the number of controls. Therefore, relying on the existing asymptotic inference theory can be problematic in empirical applications. We compare the Lasso-based inference methods to modern high-dimensional OLS-based methods and provide practical guidance.

I. Introduction

SINCE their introduction, post–double Lasso (Belloni, Chernozhukov, & Hansen, 2014) and debiased Lasso (Jeanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014) have quickly become the most popular inference methods for problems with many control variables. Given the rapidly growing theoretical1 and applied2 literature on these methods, it is crucial to take a step back and examine the performance of these procedures in empirically relevant settings and to better understand their merits and limitations relative to other alternatives.

In this paper, we study the performance of post–double Lasso and debiased Lasso from two angles. First, we develop a theory of the finite sample behavior of post–double Lasso and debiased Lasso. We show that underselection of the Lasso can lead to substantial omitted variable biases (OVBs) in finite samples. Our theoretical results on the OVBs are non-asymptotic; they complement but do not contradict the existing asymptotic theory. Second, we conduct extensive simulations and two empirical studies to investigate the performance of Lasso-based inference methods in practically relevant settings. We find that the OVBs can render the existing asymptotic approximations inaccurate and lead to size distortions.

Following the literature, we consider the following standard linear model:

\[ Y_i = D_i \alpha^* + X_i \beta^* + \eta_i, \quad i = 1, \ldots, n, \]
\[ D_i = X_i \gamma^* + \nu_i, \quad i = 1, \ldots, n. \]

Here \( Y_i \) is the outcome, \( D_i \) is the scalar treatment variable of interest, and \( X_i \) is a \((1 \times p)\)-dimensional vector of control variables. In the main text, we focus on the performance of post–double Lasso for estimating and making inferences (e.g., constructing confidence intervals) on the treatment effect \( \alpha^* \) in settings where \( p \) can be larger than or comparable to \( n \). We present results for debiased Lasso in appendix B.

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 step, the estimator of \( \alpha^*, \tilde{\alpha} \), is the OLS regression of \( Y_i \) on \( D_i \) and the union of controls selected in the two Lasso steps. OVB arises in post–double Lasso whenever the relevant controls (i.e., the controls with nonzero coefficients) are selected in neither Lasso steps, a situation we refer to as double underselection. Results that can explain when and why double underselection occurs are scarce in the literature. This paper shows theoretically that this phenomenon can even occur in simple examples with classical assumptions (e.g., normal homoskedastic errors, orthogonal designs for the relevant controls), which are often viewed as favorable to the performance of the Lasso. We prove that if the products of the absolute values of the nonzero coefficients and the variances of the controls are no greater than half the regularization parameters3 derived based on standard Lasso theory,4 Lasso fails to select these controls in both steps with high probability. This result allows us to derive the first nonasymptotic lower-bound formula in the literature for the OVB of the post–double Lasso estimator \( \tilde{\alpha} \). Our lower bound provides

3The “half the regularization parameters” type of condition on the magnitude of nonzero coefficients was independently discovered in Lahiri (2021). We are grateful to an anonymous referee for making us aware of this paper. Our proof strategies differ from the asymptotic ones in Lahiri (2021) and allow us to derive an explicit lower bound with meaningful constants for the probability of underselection for fixed \( n \), which is needed for deriving an explicit formula for the OVB lower bound. While some of the arguments in Lahiri (2021) can be made nonasymptotic, one of their core arguments for showing necessary conditions for variable selection consistency of the Lasso relies on \( n \) tending to infinity. It is not clear that such an argument can lead to an explicit lower bound with meaningful constants for the probability of underselection. In addition, our nonasymptotic argument can easily lead to asymptotic conclusions.

4Note that the existing Lasso theory requires the regularization parameter to exceed a certain threshold, which depends on the standard deviations of the noise and the covariates.
Our theoretical analysis of the OVB has important implications for inference procedures based on the post–double Lasso. Belloni et al. (2014) show that \( \sqrt{n}(\hat{\beta} - \beta^*) \) is asymptotically normal with zero mean. We show that in finite samples, the OVB lower bound can be more than twice as large as the standard deviation obtained from the asymptotic distribution in Belloni et al. (2014). This is true even when \( n \) is much larger than \( p \) and \( \beta^* \) and \( \gamma^* \) are sparse. To illustrate, assume that models (1) and (2) share the same set of \( k \) nonzero coefficients and set \( (n, p) = (14,238, 384) \) as in Angrist and Frandsen (2019, section 3), who use post–double Lasso to estimate the effect of elite colleges. The ratio of our OVB lower bound to the standard deviation in Belloni et al. (2014) is 0.27 if \( k = 1 \) and 2.4 if \( k = 10 \). This example shows that the requirement on the sparsity parameter \( k \) for the OVBs to be negligible is quite stringent. We emphasize that our findings do not contradict the existing results on the asymptotic distribution of post–double Lasso in Belloni et al. (2014). Rather, our results suggest that the OVBs can make the asymptotic zero-mean approximation of \( \sqrt{n}(\hat{\beta} - \beta^*) \) inaccurate in finite samples.

To better understand the practical implications of the OVB of post–double Lasso, we perform extensive simulations. Our simulation results can be summarized as follows. First, large OVBs are persistent across a range of empirically relevant settings and can occur even when \( n \) is large and larger than \( p \), and the sparsity parameter \( k \) is small. Second, the OVBs can lead to invalid inferences and undercoverage of confidence intervals. Third, the performance of post–double Lasso varies substantially across different popular choices of regularization parameters, and no single choice outperforms the others across all designs. While it may be tempting to choose a smaller regularization parameter than the standard recommendation in the literature to mitigate underselection, we find that this idea does not work in general and can lead to rather poor performance.

In addition to the simulations, we consider two empirical applications: the analysis of the effect of 401(k) plans on savings by Belloni et al. (2017b) and Chernozhukov et al. (2018) and the study of the racial test score gap by Fryer and Levitt (2013b). We draw samples of different sizes from the large original data and compare the subsample estimates to the estimates based on the original data.\(^5\) In both applications, we find substantial biases even when \( n \) is considerably larger than \( p \), and we document that the magnitude of the biases varies substantially depending on the regularization choice.

Given our theoretical results, simulations, and empirical evidence, a natural question is how to make statistical inferences in a reliable manner if one is concerned about OVBs. In many economic applications, \( p \) is comparable to but still smaller than \( n \). This motivates the recent development of high-dimensional OLS-based inference procedures (Cattaneo et al., 2018; D’Adamo, 2018; Johmans, 2022; Kline, Saggio, & Sølvsten, 2020). These methods are based on OLS regressions with all controls and rely on novel variance estimators that are robust to the inclusion of many controls (unlike conventional variance estimators). Based on extensive simulations, we find that OLS with standard errors proposed by Cattaneo et al. (2018) demonstrates excellent coverage accuracy across all our simulation designs. Another advantage of OLS-based methods over Lasso-based inference methods is that the former do not rely on any sparsity assumptions. This is important because sparsity assumptions may not be satisfied in applications, and as this paper shows, the OVBs of Lasso-based inference procedures can be substantial even when \( k \) is small and \( n \) is large and larger than \( p \). However, OLS yields somewhat wider confidence intervals than the Lasso-based inference methods, suggesting a trade-off between coverage accuracy and the length of the confidence intervals.

\(^5\)For example, Kolesár and Rothe (2018) use a similar exercise to illustrate the issues with discrete running variables in regression discontinuity designs.
Our analyses suggest two main recommendations concerning the use of post–double Lasso in empirical studies. First, if the estimates of \( \alpha^* \) are robust to increasing the recommended regularization parameters in both Lasso steps, this suggests that either the OVBs are negligible (regime b) or underselection is unlikely (regime c). In either case, post–double Lasso is a reliable and efficient method. Otherwise, modern high-dimensional OLS-based inference methods constitute a possible alternative when \( p \) is smaller than \( n \). Second, our findings highlight the importance of augmenting the final OLS regression in post–double Lasso with control variables motivated by economic theory and prior knowledge, as suggested by Belloni et al. (2014).

II. Lasso and Post–Double Lasso

A. The Lasso

Consider the following linear regression model,

\[
Y_i = X_i \theta^* + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}
\]

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

The Lasso estimator of \( \theta^* \), first proposed by Tibshirani (1996), 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 regularization parameter. Let \( \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n) \) and \( X \) be a fixed design matrix with normalized columns \((n^{-1} \sum_{i=1}^n X_{ij}^2 = 1 \) for all \( j = 1, \ldots, p \)). In this example, Bickel, Ritov, and Tsybakov (2009) set \( \lambda = 2 \sigma \sqrt{2n^{-1} (1 + \tau)} \log p \) (where \( \tau > 0 \)) to establish upper bounds on \( \sum_{j=1}^p (\hat{\theta}_j - \theta^*_j)^2 \) with a high-probability guarantee. To establish perfect selection, Wainwright (2009) sets \( \lambda \) proportional to \( \sigma \phi^{-1} \sqrt{\log p} / n \), where \( \phi \in (0, 1) \) is a measure of correlation between the covariates with nonzero coefficients and those with zero coefficients.

Besides the classical choices in Bickel et al. (2009) and Wainwright (2009), other choices of \( \lambda \) are available in the literature. For instance, Belloni, Chen et al. (2012) and Belloni, Chernozhukov et al. (2016) propose choices that accommodate heteroskedastic and clustered errors. The regularization choice of Belloni et al. (2012), which is recommended by Belloni et al. (2014) for post–double Lasso, is based on the following Lasso program:

\[
\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \theta)^2 + \frac{\lambda}{n} \sum_{j=1}^p |\hat{\theta}_j|, \tag{5}
\]

where \( \{\hat{\theta}_1, \ldots, \hat{\theta}_p\} \) are penalty loadings obtained using the iterative algorithm developed in Belloni et al. (2012).

Finally, a very popular practical approach for choosing \( \lambda \) is cross-validation; see, for example, Homrighausen and McDonald (2013, 2014) and Chetverikov, Liao, and Chernozhukov (2021) for theoretical results on cross-validated Lasso.

B. Post–Double Lasso

The model (1)–(2) implies the following reduced-form model for \( Y_i \),

\[
Y_i = X_i \pi^* + u_i, \tag{6}
\]

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

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

\[
\hat{\pi} \in \arg \min_{\pi \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (Y_i - X_i \pi)^2 + \lambda_1 \sum_{j=1}^p |\pi_j|, \tag{7}
\]

\[
\hat{\gamma} \in \arg \min_{\gamma \in \mathbb{R}^c} \frac{1}{2n} \sum_{i=1}^n (D_i - X_i \gamma)^2 + \lambda_2 \sum_{j=1}^p |\gamma_j|. \tag{8}
\]

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

\[
(\hat{\alpha}, \hat{\beta}) \in \arg \min_{\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^c} \frac{1}{2n} \sum_{i=1}^n (Y_i - D_i \alpha - X_i \beta)^2 \quad \text{s.t.} \quad \beta_j = 0 \ \forall j \notin \{\hat{I}_1 \cup \hat{I}_2\}, \tag{9}
\]

where \( \hat{I}_1 = \text{supp}(\hat{\pi}) = \{j : \hat{\pi}_j \neq 0\} \) and \( \hat{I}_2 = \text{supp}(\hat{\gamma}) = \{j : \hat{\gamma}_j \neq 0\} \).

III. Numerical Example

This section presents a simple numerical example illustrating the OVB of post–double Lasso. All computations were performed in Matlab (Matlab, 2020). The Lasso is implemented using the built-in function lasso. We consider a simple but classical setting that is often considered favorable to the performance of the Lasso. The data are simulated according to the structural model in equations (1) and (2), where \( X_i \overset{iid}{\sim} \mathcal{N}(0_p, I_p) \), \( \eta_i \overset{iid}{\sim} \mathcal{N}(0, 1) \), and \( v_i \overset{iid}{\sim} \mathcal{N}(0, 1) \) are independent of each other. Our object of interest is \( \alpha^* \). We set \( n = 500, p = 200, \alpha^* = 0 \), and consider a sparse setting where \( \beta_j^* = \gamma_j^* = c \cdot 1 \) \( (j \leq k) \) for \( j = 1, \ldots, p \) and \( k = 5 \). Following the simulation exercise in Belloni et al. (2014), we vary the population \( R^2 \)’s in equations (2) and (6) by varying the magnitude of the nonzero coefficients \( c \). We employ the regularization parameter choice by Bickel et al. (2009).

Figure 2 displays the finite sample distribution of post–double Lasso for different values of \( R^2 \). For comparison, we
The gray histograms show the finite sample distributions, and the black curves show the densities of the oracle estimators.

The simple numerical example in this section shows that post–double Lasso can suffer from OVBs when the two Lasso steps do not select all relevant controls. The magnitude of the coefficients corresponding to the omitted controls can be large enough that the OVB shifts the location of the finite sample distribution far from the true value $\alpha^* = 0$. The issue documented here is not a “small sample” phenomenon but persists even in large sample settings (see appendix C.1).

IV. Theoretical Analysis

This section provides a theoretical analysis of the OVB of post–double Lasso. Our goal here is to demonstrate that even in simple examples with classical assumptions (e.g., normal homoskedastic errors, orthogonal designs for the relevant controls), which are often viewed favorable to the performance of Lasso, the finite sample OVBs of post–double Lasso can be substantial relative to the standard deviation provided in the literature. We first establish a new necessary result for the Lasso’s inclusion and then derive lower and upper bounds on the OVBs of post–double Lasso. These results are derived for fixed $(n, p, k)$ and are also valid when $(k \log p) / n \to 0$ or $(k \log p) / n \to \infty$. As it will become clear in the following, $p$ needs to be large enough for our results to be informative. Without loss of generality, we normalize the matrix $X$ such that $(X_j'X_j) / n = 1$ for all $j = 1, \ldots, p$. We focus on fixed designs (of $X$) to highlight the essence of the problem (see appendix D for an extension to random designs).
Here, we present the notation to be used in the theoretical analysis. Let $1_m$ denote the $m$-dimensional (column) vector of “1”s and $0_m$ is defined similarly. The $\ell_1$-norm of a vector $v \in \mathbb{R}^m$ is denoted by $|v|_1 := \sum_{i=1}^m |v_i|$, and the $\ell_\infty$-norm of a vector $v \in \mathbb{R}^m$ is denoted by $|v|_\infty := \max_{i=1,\ldots,m} |v_i|$. 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 subvector (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 $v_j = 0$. Given a set $K$, let $\text{card}(K)$ denote the cardinality of $K$. We denote max $\{a, b\}$ by $a \vee b$ and min $\{a, b\}$ by $a \wedge b$.

### A. Stronger Necessary Results on the Lasso’s Inclusion

Post–double Lasso exhibits OVBs whenever the relevant controls are selected in neither program (7) nor program (8). To the best of our knowledge, there are no formal results strong enough to show that, with high probability, Lasso can fail to select the relevant controls in both steps. Therefore, we first establish a new necessary result for the single Lasso’s inclusion in Lemma 1. To derive this result, we consider the following classical assumptions, which are often viewed favorable to the performance of the Lasso.

**Assumption 1.** In terms of model (3), suppose: (i) $K = \{j : \hat{\theta}_j \neq 0\} \neq \emptyset$ and card $(K) = k \leq (n \wedge p)$; (ii) $X_K^T X_K$ is a diagonal matrix; (iii) $\| X_K^T X_K (X_K^T X_K)^{-1} \|_\infty = 1 - \phi$ for some $\phi \in (0, 1)$, where $K^c$ is the complement of $K$.

Known as the incoherence condition due to Wainwright (2009), part iii in assumption 1 is needed for the exclusion of the irrelevant controls. Note that if the columns in $X_K$ are orthogonal to the columns in $X_K$ (but within $X_K$, the columns need not be orthogonal to each other), then $\phi = 1$. Obviously a special case of this is when the entire $X$ consists of mutually orthogonal columns (which is possible if $n \geq p$). To provide some intuition for assumption 1(iii), consider the simple case where $k = 1$ and $K = \{1\}$, $X$ is centered (such that $n^{-1} \sum_{i=1}^n X_{i1} = 0$), and the columns in $X$ are normalized such that the standard deviations of $X_1$ and $X_j$ (for any $j \in \{2, 3, \ldots, p\}$) are identical. Then, $1 - \phi$ is simply the maximum of the absolute (sample) correlations between $X_1$ and each of the $X_j$s with $j \in \{2, 3, \ldots, p\}$.

**Lemma 1** (Necessary result on the Lasso’s inclusion). In model (3), suppose the $\varepsilon_i$s are independent over $i = 1, \ldots, n$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $\sigma \in (0, \infty)$. Let assumption 1 hold. We solve the Lasso program (4) with

$$
\lambda \geq \frac{2\sigma}{\phi} \sqrt{\frac{2(1 + \tau) \log p}{n}} \quad (10)
$$

where $\tau > 0$. Let $E_1$ denote the event that $\text{sgn}(\hat{\theta}_j) = -\text{sgn}(\theta^*_j)$ for at least one $j \in K$, and $E_2$ denote the event that $\text{sgn}(\hat{\theta}_l) = \text{sgn}(\theta^*_l)$ for at least one $l \in K$ with

$$
|\theta^*_l| \leq \frac{\lambda}{2}. \quad(11)
$$

Then we have

$$
\mathbb{P}(E_1 \cap \mathcal{E}) = \mathbb{P}(E_2 \cap \mathcal{E}) = 0,
$$

where $\mathcal{E} = \{ |n^{-1}X^T \varepsilon|_\infty \leq \sigma\phi^{-1} \sqrt{2n^{-1}(1 + \tau) \log p} \}$ and $\mathbb{P}(\mathcal{E}) \geq 1 - 1/p^\tau$.

If inequality (11) holds for all $l \in K$, we have

$$
\mathbb{P}(\hat{\theta}_0 = 0_p) \geq 1 - \frac{1}{p^\tau}. \quad (13)
$$

Lemma 1 shows that for large enough $p$, Lasso fails to select any of the relevant covariates with high probability if inequality (11) holds for all $l \in K$. If such conditions hold with respect to both models (1) and (2), then lemma 1 implies that the relevant controls are selected in neither program (7) nor (8) with probability at least $1 - 2/p^\tau$ (see figures 3c and 3d for an illustration).

We rewrite inequality (11) as $|\theta^*_l| = 2^{-1} |a| \lambda$ with $|a| \in (0, 1]$. Assume that $|a| \phi^{-1} \sigma$ is bounded from above and away from 0; moreover, $\lambda$ satisfies inequality (10) and scales as $\sqrt{(\log p) / n}$. These conditions imply that $|\theta^*_l| \sim \sqrt{\lambda / n}$ in the classical asymptotic framework where $n \rightarrow \infty$ and $p$ is fixed. This regime of $\theta^*_l$ is exactly where classical model selection procedures struggle to distinguish a coefficient from 0 in low-dimensional settings.

In section I, we discussed the relationship of our result to that in Lahiri (2021). It is also interesting to compare lemma 1 with the results in Wainwright (2009). Note that equation (12) implies $\mathbb{P}(\hat{\theta}_l = 0) \leq 1/p^\tau$ for any $l \in K$ subject to inequality (11). In comparison, Wainwright (2009) shows that whenever $\theta^*_l \in (\lambda \text{sgn}(\theta^*_l), 0)$ or $\theta^*_l \in (0, \lambda \text{sgn}(\theta^*_l))$ for some $l \in K$,

$$
\mathbb{P}\left[\text{sgn}(\hat{\theta}_K) = \text{sgn}(\theta^*_K)\right] \leq \frac{1}{2}. \quad (14)
$$

Constant bounds in the form of inequality (14) cannot explain that, with high probability, Lasso fails to select the relevant covariates in both programs (7) and (8) when $p$ is sufficiently large.

**Remark 1.** As the choices of regularization parameters used in the vast majority of literature (e.g., Bickel et al., 2009; Wainwright, 2009; Belloni, Chen et al., 2012; Belloni, Chernozhukov et al., 2014; Belloni & Chernozhukov, 2013), the choice of $\lambda$ in lemma 1 is derived from the principle that $\lambda$ should be no smaller than $2 \max_{j=1,\ldots,p} |n^{-1}X_j^T \varepsilon|$ with high probability. In particular, our choice for $\lambda$ takes the form
that in Wainwright (2009), but ours involves a sharper universal constant. Choosing regularization parameters in this form ensures the exclusion of irrelevant controls. In addition, our choice for \( \lambda \) has a scaling that can be achieved by the regularization parameters in Belloni et al. (2012), Belloni, Chernozhukov et al. (2014), and Belloni and Chernozhukov (2013) and coincides with that in Bickel et al. (2009) when the columns in \( X_K^\perp \) are orthogonal to the columns in \( X_K \) (but within \( X_K^\perp \), the columns need not be orthogonal to each other).

B. Lower Bounds on the OVBs

In this section, we apply lemma 1 to derive lower bounds on the OVB of post–double Lasso. We consider the structural models (1) and (2), which can be written in matrix notation as

\[ Y = D\alpha^* + X\beta^* + \eta, \quad (15) \]
\[ D = X\gamma^* + v. \quad (16) \]

In matrix notation, the reduced form, model (6), becomes

\[ Y = X\pi^* + u, \quad (17) \]

where \( \pi^* = \gamma^*\alpha^* + \beta^* \) and \( u = \eta + \alpha^*v \). We make the following assumptions about models (15) and (16).

**Assumption 2.** (i) The error terms \( \eta \) and \( v \) consist of independent entries drawn from \( \mathcal{N}(0, \sigma^2_\eta) \) and \( \mathcal{N}(0, \sigma^2_v) \), respectively, where \( \eta \) and \( v \) are independent of each other; (ii) the data are centered: \( D = n^{-1}\sum_{i=1}^n D_i = 0, X = \{n^{-1}\sum_{i=1}^n X_{ij}\}^T_{j=1} = 0_p, \) and \( Y = n^{-1}\sum_{i=1}^n Y_i = 0; \) and (iii) \( K = \{j : \beta_j^* \neq 0\} = \{j : \gamma_j^* \neq 0\} \neq \emptyset \) and card \((K) = k \leq (n \wedge p)\).

Proposition 1 derives a lower-bound formula for the OVB of post–double Lasso concerning the case where \( \alpha^* = 0 \).

**Proposition 1 (OVB lower bound).** Suppose \( \alpha^* = 0 \). Let assumption 1(ii)–(iii) and assumption 2 hold; \( \lambda_1 = 2\phi^{-1}\sigma_\eta \sqrt{2n^{-1}(1 + \tau)\log p} \) and \( \lambda_2 = 2\phi^{-1}\sigma_\eta \sqrt{2n^{-1}(1 + \tau)\log p}; \) for all \( j \in K \) and \( |a|, |b| \in (0, 1] \).

\[ \begin{align*}
\beta_j^* &= a\phi^{-1}\sigma_\eta \sqrt{2(1 + \tau)\log p} / n, \\
\gamma_j^* &= b\phi^{-1}\sigma_\eta \sqrt{2(1 + \tau)\log p} / n.\
\end{align*} \quad (18) \]

In terms of \( \hat{\alpha} \) obtained from program (9), we have

\[ |E(\hat{\alpha} - \alpha^*|\mathcal{M})| \geq \max_{r \in (0,1]} \text{max} \left\{ T_1(r) T_2(r) \right\} := \text{OVB} \]

where

\[ T_1(r) = \frac{(1 + \tau)|ab|\phi^{-2}\sigma_\eta \sqrt{k\log p / n}}{4(1 + \tau)\phi^{-2}b^2\sigma_\eta \sqrt{k\log p / n} + (1 + r)\sigma_v}, \]
\[ T_2(r) = 1 - k \exp\left(-\frac{b^2(1 + \tau)\log p}{4\phi^2} - \frac{1 - p}{p^*} - \exp\left(-\frac{nr^2}{8}\right)\right), \]

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

**Remark 2.** In our theoretical results, we implicitly assume \( p \) is sufficiently large such that \( 1 - k \exp\left(-\frac{1}{2}\right)^{-1} b^2(1 + \tau)\log p - 2/p^* > 0 \). Indeed, probabilities in such a form are often referred to as the “high-probability” guarantees in the literature of (nonasymptotic) high-dimensional statistics concerning large \( p \) and small enough \( k \). Recalling the definitions of \( \hat{I}_1 \) and \( \hat{I}_2 \) in section IIb, the event \( \mathcal{M} \) is the intersection of \( \{\hat{I}_1 = \hat{I}_2 = \emptyset\} \) and an additional event \( \mathcal{E}_r = \{n^{-1}X_K^Tv_{|\infty} \leq t^*, t^* = 4^{-1}|b|_2\}. \) The event \( \{\hat{I}_1 = \hat{I}_2 = \emptyset\} \) occurs with probability at least \( 1 - 2/p^* \), and the event \( \mathcal{E}_r \) occurs with probability at least \( 1 - k \exp\left(-\frac{1}{2}\right)^{-1} b^2(1 + \tau)\log p \). The additional event \( \mathcal{E}_r \) is needed for us to derive a non-trivial lower bound. In particular, with probability at most \( k \exp\left(-\frac{1}{2}\right)^{-1} b^2(1 + \tau)\log p \), we have \( |n^{-1}X_K^Tv_{|\infty} \geq t^* \), and on this event, the lower bound in proposition 1 can be negative, which is uninformative for the absolute value of OVBs.

Let us compare the nonasymptotic lower bound in proposition 1 to the implications of the existing asymptotic results for the bias of post–double Lasso. If \( \sigma_\eta \) is bounded away from 0 and \( \sigma_v \) is bounded from above, the existing theory would imply that the biases of post–double Lasso are bounded from above by constant \( (k \log p) / n \), irrespective of whether Lasso fails to select the relevant controls, and how small \( |a| \) and \( |b| \) are. The (positive) constant does not depend on \( (n, p, k) \) and is implicit about the roles \( \beta_K^* \) and \( \gamma_K^* \) play. Moreover, this constant bears little meaning in the asymptotic framework, which simply assumes \( (k \log p) / \sqrt{n} \rightarrow 0 \) among other sufficient conditions. (The existing theoretical framework makes it difficult to derive an informative constant, and to our knowledge, the literature provides no such derivation.) The asymptotic upper–bound constant \( (k \log p) / n \) does not distinguish cases that vary in \( (\beta_K^*, \gamma_K^*, \alpha^*) \). By contrast, our lower-bound analyses are informative about whether the upper bound can be attained by the magnitude of the OVBs and provide explicit constants. These features of our analysis are crucial for understanding the finite sample limitations of post-double Lasso. In view of proposition 1, OVB is not a simple linear function of \( (k \log p) / n \) in general but roughly linear in \( (k \log p) / n \) when \( \phi^{-2}n^{-1}k \log p = o(1) \) and \( k \exp\left(-\frac{1}{2}\right)^{-1} b^2(1 + \tau)\log p + 2/p^* = o(1) \).
C. Key Takeaways of Our Theoretical Results

The finite sample behavior of post–double Lasso can be characterized by three regimes: (i) nonnegligible OVBs, (ii) negligible OVBs, and (iii) the absence of OVBs.

Regime i (nonnegligible OVB). When double underselection occurs with high probability and \((k \log p) / \sqrt{n}\) is not small enough, according to proposition 1, the OVB lower bound can be substantial compared to the standard deviation obtained from the asymptotic distribution in Belloni et al. (2014). To gauge the magnitude of the OVB and explain why the confidence intervals proposed in the literature can exhibit undercoverage, it is instructive to compare OVB with \(\sigma_a = n^{-1/2} (\sigma_a / \sigma_n)\), the standard deviation (of \(\hat{a}\)) obtained from the asymptotic distribution in Belloni et al. (2014). Let us consider an example with \(n = 14,238, p = 384\) (as in Angrist & Frandsen, 2019), \(a = b = 1, \sigma_a = \sigma_0 = 1, \tau = 0.5, \) and \(p = 0.5\). If \(|\beta_j| = 0.07\) and \(|\gamma_j| = 0.07\) for all \(j \in K\), then \(\text{OVB} / \sigma_a \approx 0.27\) when \(k = 1\) and \(\text{OVB} / \sigma_a \approx 2.4\) when \(k = 10\); see figures 2c and 3c for an illustration of regime i. The OVBs can also be nonnegligible when some but not all relevant controls are selected; see figures 2b and 3b. These results suggest that, nonasymptotically, post–double Lasso cannot avoid the “post-selection inference issues” raised in a series of papers by Leeb and Pötscher (2005, 2008, 2017).

Regime ii (negligible OVB). By lemma 1 and equation (18), \(|\hat{\beta} - \beta^*| = 2^{-1} |a| k \lambda_1\) and \(|\hat{\gamma} - \gamma^*| = 2^{-1} |b| k \lambda_2\) with probability at least \(1 - 2/p^5\). If \(\sigma_a\) is bounded away from 0, \(\sigma_a\) is bounded from above, and \(|a| = |b| = \sigma_0 (1)\), by a similar argument as in Belloni et al. (2014), we can show that \(\sqrt{n} (\hat{a} - \alpha^*)\) is approximately normal and centered at 0, even if \((k \log p) / \sqrt{n}\) is bounded away from 0 and scales as a constant. Holding other factors constant, the magnitude of OVBs decreases as \(\hat{\beta}^*_k\) and \(\gamma_k^*\) decrease (i.e., as \(|a|\) and \(|b|\) decrease). As \(|a|\) and \(|b|\) become very small, the relevant controls become essentially irrelevant. Figures 2d and 3d provide an illustration of regime ii.

Regime iii (absence of OVB). All relevant controls will be selected when the magnitude of their coefficients is large enough. Specifically, for all \(j \in K\), \(a > 3, b > 3\), if

\[
\begin{align*}
|\beta_j^*| &= a \phi^{-1} \sigma_n \sqrt{\frac{2(1 + \tau) \log p}{n}} \\
|\gamma_j^*| &= b \phi^{-1} \sigma_n \sqrt{\frac{2(1 + \tau) \log p}{n}},
\end{align*}
\]

then \(\mathbb{P}[\text{supp} (\hat{a}) = \text{supp} (\pi^* \gamma^* )] \geq 1 - 1/p^3\) or \(\mathbb{P}[\text{supp} (\hat{\gamma}) = \text{supp} (\gamma^* )] \geq 1 - 1/p^3\) by standard arguments (Wainwright, 2019). As a result, \(\mathbb{P}(\{\hat{I}_1 \cup \hat{I}_2 = K\}) \geq 1 - 1/p^3\) (where \(\hat{I}_1\) and \(\hat{I}_2\) are defined in section IIB); that is, the final OLS step, program (9), includes all the relevant controls with high probability. By similar argument as in appendix A.2, on the high-probability event \(\{\hat{I}_1 \cup \hat{I}_2 = K\}\), the OVB of \(\hat{a}\) is 0. Figures 2a and 3a provide an illustration of regime iii.

This “triple-regime” characterization suggests that one may assess the robustness of post–double Lasso by increasing the penalty level. If increasing \(\lambda_1\) and \(\lambda_2\) yields similar estimates \(\hat{a}\), then the underlying model could be in the regime where either the OVBs are negligible (regime ii) or underselection in both Lasso steps is unlikely (regime iii). By contrast, under regime i, the performance of post–double Lasso can be quite sensitive to an increase of \(\lambda_1\) and \(\lambda_2\). The rationale behind this heuristic lies in that the final step of post–double Lasso, program (9), is simply an OLS regression of \(Y\) on \(D\) and the union of selected controls from programs (7) and (8). A natural question is by how much \(\lambda_1\) and \(\lambda_2\) should be increased for the robustness checks. For the regularization choice proposed in Belloni et al. (2014), we will show in the simulations of section V that an increase by 50% works well in practice.

Finally, our theoretical results have interesting implications for the comparison between post–double Lasso and post–single Lasso, where the latter only relies on one Lasso step to select the relevant controls. Note that the magnitude of the OVB of the post–single Lasso estimator of \(\alpha^*\) also falls into three regimes: (i) nonnegligible OVB, (ii) negligible OVB, and (iii) the absence of OVB. Thus, qualitatively, the OVBs of post–single Lasso and post–double Lasso have a similar behavior in finite samples. However, quantitatively, the magnitude of the OVBs can be much larger for the post–single Lasso than for the post–double Lasso, as illustrated by the following example.

If \(\beta_j = a \phi^{-1} \sigma_n \sqrt{2n^{-1}(1 + \tau) \log p}\) with \(|a| \in (0, 1]\), and \(|\gamma_j| \geq 1\), then the argument for showing proposition 1 in appendix A.2 implies that the OVB lower-bound scales roughly as \(|a| k (\log p) / n\) for the post–single Lasso estimator of \(\alpha^*\) (and note that \(|a| k (\log p) / n| / (1 / \sqrt{n}) = |a| k \sqrt{\log p}\).

D. Additional Theoretical Results

In this section, we briefly summarize the additional theoretical results that are provided in the appendix. First, we also consider cases where \(\alpha^* \neq 0\). The conditions required to derive the explicit formula are difficult to interpret when \(\alpha^* \neq 0\). However, it is possible to provide easy-to-interpret scaling results (without explicit constants) for cases where \(\alpha^* \neq 0\). Roughly, the scaling of our OVB lower bound can be as large as

\[
\left(\frac{\sigma_{\pi} \sigma_n}{\sigma_v \alpha^*} \right) \frac{k \log p}{n} \text{ when } \frac{k \log p}{n} \to 0
\]

and

\[
\frac{\sigma_n}{\sigma_v} \frac{|\alpha^*|}{\alpha^*} \text{ when } \frac{k \log p}{n} \to \infty.
\]
These results reveal an interesting feature of the post–double Lasso. The scaling of the OVB lower bound depends on $|\alpha^*|$ when the relevant controls are not selected. This is because the error $u$ in the reduced-form equation (17) involves $\alpha^*$ such that the choice of $\lambda_1$ in program (7) depends on $|\alpha^*|$ via the variance of $u$. By contrast, it is well known that the OVB of OLS does not depend on $|\alpha^*|$ when relevant controls are omitted. (Interested readers are referred to propositions A.1 and A.2 in appendix A.3 for details.)

Second, we also provide upper bounds on the OVB. We have seen in expression (20) that the lower bound on the OVBs scales as $(\sigma_n/\sigma_0) \lor |\alpha^*|$ when $(k \log p)/n \to \infty$. Interestingly enough, we can also show that the upper bound on the OVB scales as in expression (20) despite $(k \log p)/n \to \infty$ and the Lasso being inconsistent in the sense

\[
\sqrt{n^{-1} \sum_{i=1}^{n} (X_i \hat{\pi} - X_i \pi^*)^2} \to \infty, \quad \sqrt{n^{-1} \sum_{i=1}^{n} (X_i \hat{\gamma} - X_i \gamma^*)^2} \to \infty \quad \text{with high probability.} \quad \text{(Interested readers are referred to propositions A.3 and A.4 in appendix A.6 for details.)}
\]

V. Simulations and Empirical Evidence

To better understand the practical implications of the OVB of post–double Lasso, in this section, we present the results from simulations and two empirical applications with widely used regularization choices available in standard software packages. The analyses were carried out using R (R Core Team, 2021), and Stata (StataCorp, 2021).

A. Simulations

In this section, we present simulation evidence on the performance of post–double Lasso with three choices of the regularization parameter: (i) the heteroskedasticity-robust proposal of Belloni, Chen et al. (2012) and Belloni, Chernozhukov et al. (2014) ($\lambda_{BCCH}$) implemented using the R-package hdm with the double selection option (Chernozhukov et al., 2016), (ii) the regularization parameter with the minimum cross-validated error ($\lambda_{min}$) implemented using the R-package glmnet (Friedman et al., 2010), and (iii) the regularization parameter corresponding to the minimum plus 1 standard deviation cross-validated error ($\lambda_{1se}$) also implemented using glmnet. We use the same type of regularization parameter choice in both Lasso steps.

We simulate data according to the DGP of section III. To illustrate the role of the sample size $n$ and the sparsity parameter $k$, we consider (i) $(n, p, k) = (500, 200, 5)$, (ii) $(n, p, k) = (1000, 200, 5)$, and (iii) $(n, p, k) = (500, 200, 10)$. We show results for $R^2 \in [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]$ based on 1,000 simulation repetitions. Appendix C presents additional simulation evidence, where we vary $n$, the distribution of $(X_i, \eta_i, \nu_i)$, the true value $\alpha^*$, and also consider a heteroskedastic DGP.

We start by investigating the selection performance of the two Lasso steps of post–double Lasso. Figure 4a displays the average number of selected controls (i.e., the cardinality of $I_1 \cup I_2$ in program (9)) as a function of $R^2$. Lasso with $\lambda_{BCCH}$ selects the lowest number of controls. Choosing $\lambda_{1se}$ leads to a somewhat higher number of selected controls and results in moderate overselection for larger values of $R^2$. Lasso with $\lambda_{min}$ selects the highest number of controls and exhibits substantial overselection. Panel b shows the corresponding average numbers of selected relevant controls.

Figure 5 presents evidence on the bias of post–double Lasso. To make the results easier to interpret, we report the ratio of the bias to the empirical standard deviation. Post–double Lasso with $\lambda_{BCCH}$ can exhibit biases that are more than two times larger than the standard deviation when $(n, p, k) = (500, 200, 10)$. The bias can still be comparable to the standard deviation when $(n, p, k) = (1000, 200, 5)$. Consistent with our theoretical discussions in sections IVB and IVC, the relationship between $R^2$ and the ratio of bias to standard deviation is nonmonotonic: it is increasing for small $R^2$ and decreasing for larger $R^2$. The bias is somewhat smaller for $\lambda = \lambda_{1se}$. Setting $\lambda = \lambda_{min}$ yields the smallest ratio of bias to standard deviation. Finally, we note that when $R^2$ is large enough such that there is no underselection, post–double Lasso performs well and is approximately unbiased for all regularization parameters.

The additional simulation evidence reported in appendix C confirms these results but 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 coefficients and the error in the reduced-form equation (6). Moreover, we show that while choosing $\lambda = \lambda_{min}$ works well when $\alpha^* = 0$, this choice can yield poor performances when $\alpha^* \neq 0$ (see figure C.2). A similar phenomenon arises when using $0.5\lambda_{BCCH}$ instead of $\lambda_{BCCH}$: this choice works well when $\alpha^* = 0$ (see figure 7 below) but yields biases when $\alpha^* \neq 0$. We found that under our DGPs, this is related to the fact that when $\alpha^* \neq 0$, models (2) and (6) differ with respect to the underlying coefficients and errors, which leads to differences in the (over-)selection behavior of the post-double Lasso. Thus, there is no simple recommendation for how to choose the regularization parameters in practice.

The substantive performance differences between the three regularization choices suggest that post–double Lasso is sensitive to the penalty levels in the intermediate case where $R^2$ is small enough so that underselection occurs but large enough to cause substantial OVBs. To further investigate this issue, we compare the results for $\lambda_{BCCH}$, $0.5\lambda_{BCCH}$, and $1.5\lambda_{BCCH}$. Figure 6 displays the average numbers of all selected controls (relevant or not) and selected relevant controls in both Lasso steps. The differences in the selection performance are substantial. Lasso with $0.5\lambda_{BCCH}$ overselects for all $n$ and $R^2$, while Lasso with $1.5\lambda_{BCCH}$ underselects unless $R^2$ and $n$ are large and $k = 5$. The differences get smaller as $n$ increases and larger as $k$ increases.

Figure 7 displays the ratio of bias to standard deviation. Choosing $0.5\lambda_{BCCH}$ yields small biases relative to the standard deviations for all $R^2$. By contrast, choosing $1.5\lambda_{BCCH}$ yields biases that can be more than six times larger than the
standard deviations when \((n, p, k) = (500, 200, 10)\) and still substantial when \((n, p, k) = (1,000, 200, 5)\). For very small and large values of \(R^2\), post–double Lasso is less sensitive to the penalty level. In section VII, we discuss how to interpret and use robustness checks with respect to the regularization parameters in empirical applications.

In sum, our simulation evidence shows that underselection can lead to large biases relative to the standard deviations, the performance of post–double Lasso can be very sensitive to the choice of regularization parameters, and there is no simple recommendation for how to choose the regularization parameters in practice.
B. Empirical Evidence

The effect of 401(k) plans on total wealth. We revisit the analysis of the causal effect of eligibility for 401(k) plans ($D$) on total wealth ($Y$). We use the data on $n = 9,915$ households from the 1991 SIPP (Belloni et al., 2017a) analyzed by Belloni et al. (2017b) and Chernozhukov et al. (2018) with holds from 1995, 1998) and Benjamin (2003). Other studies have used 401(k) eligibility to instrument for the actual 401(k) participation status (e.g., Abadie, 2003; Chernozhukov & Hansen, 2004; Belloni et al., 2017b; Wüthrich, 2019).

---

![Figure 6](image1.png)

**Figure 6.**—Average number of selected controls: Sensitivity to penalty level

(a) All controls

- $n=500, p=200, k=5$
- $n=1000, p=200, k=5$
- $n=500, p=200, k=10$

![Figure 7](image2.png)

**Figure 7.**—Ratio of bias to standard deviation: Sensitivity to penalty level

(b) Relevant controls

- $n=500, p=200, k=5$
- $n=1000, p=200, k=5$
- $n=500, p=200, k=10$
1. Two-way interactions (TWI) specification. We use the same set of low-dimensional control variables as in Benjmin (2003) and Chernozhukov and Hansen (2004): seven income dummies, five age dummies, family size, four education dummies, and dummies for marital status, two-earner status, defined-benefit pension status, individual retirement account (IRA) participation status, and homeownership. Following common empirical practice, we augment this baseline specification with all two-way interactions. After removing collinear columns, there are $p = 167$ control variables.

2. Quadratic spline and interactions (QSI) specification. This is the “quadratic spline plus interactions specification” of Belloni et al. (2017b, p. 265). It contains dummies for marital status, two-earner status, defined-benefit pension status, IRA participation status, and homeownership, second-order polynomials in family size and education, a third-order polynomial in age, a quadratic spline in income with six breakpoints, as well as interactions of all the non-income variables with each term in the income spline. After removing collinear columns there are $p = 272$ control variables.

Table 1 presents post–double Lasso estimates based on the whole sample with $\lambda_{BCCH}$, $0.5\lambda_{BCCH}$, and $1.5\lambda_{BCCH}$. For comparison, we also report OLS estimates with and without controls. For both specifications, the results are qualitatively similar across the different regularization choices and similar to OLS with all controls. This is possible as $n$ is much larger than $p$. Nevertheless, there are some nonnegligible quantitative differences between the point estimates. A comparison to OLS without control variables shows that omitting controls can yield substantial OVBs in this application.

To investigate the impact of underselection, we perform the following exercise. We draw random subsamples of size $n_s \in \{200, 400, 800, 1,600\}$ with replacement from the original data set. Based on each subsample, we estimate $\alpha^*$ using post–double Lasso with $\lambda_{BCCH}$, $0.5\lambda_{BCCH}$, and $1.5\lambda_{BCCH}$ and compute the bias as the difference between the average sub-sample estimate and the point estimate based on the original data with the same type of regularization choice in table 1. The results are based on 1,000 simulation repetitions.

Figure 8 displays the bias and the ratio of bias to standard deviation for both specifications. We find that post–double Lasso can exhibit large finite sample biases. The biases under the QSI specification tend to be smaller (in absolute value) than the biases under the TWI specification. Interestingly, the ratio of bias to standard deviation may not be monotonically decreasing in $n_s$ (in absolute value) due to the standard deviation decaying faster than the bias. Finally, we find that post–double Lasso can be very sensitive to the penalty level.

Table 2 shows the results for post–double Lasso with $\lambda_{BCCH}$, $0.5\lambda_{BCCH}$, and $1.5\lambda_{BCCH}$, as well as OLS with and without controls based on the whole sample. Since $n = 30,002$ is much larger than $p = 78$, all methods except for OLS without controls yield similar results.

To investigate the impact of underselection, we draw random subsamples of size $n_s \in \{200, 400, 800, 1,600\}$ with replacement from the original data set. In each sample, we estimate $\alpha^*$ using post–double Lasso with $\lambda_{BCCH}$, $0.5\lambda_{BCCH}$, and $1.5\lambda_{BCCH}$ and compute the bias as the difference between the average estimate based on the subsamples and the estimate based on the original data with the same type of regularization choice. The results are based on 1,000 simulation repetitions.

Figure 9 displays the bias and the ratio of bias to standard deviation. While the magnitude of the bias is decreasing in $n_s$, it can be substantial and larger than the standard deviation when $n_s$ is small. Moreover, the performance of post–double Lasso is very sensitive to the choice of the regularization parameters. With $0.5\lambda_{BCCH}$, post–double Lasso is approximately unbiased for all $n_s$ whereas with $1.5\lambda_{BCCH}$, the bias is comparable to the standard deviation even when $n_s = 1,600$. Racial differences in the mental ability of children. We revisit Fryer and Levitt’s (2013b) analysis of the racial differences in the mental ability of young children based on data from the U.S. Collaborative Perinatal Project (Fryer & Levitt, 2013a). As in the reanalysis of Chernozhukov et al. (2020), we restrict the sample to Black and White children so that our final sample includes $n = 30,002$ observations. We focus on the standardized test score in the Wechsler Intelligence Test at the age of 7 as our outcome variable ($Y$). The variable of interest ($D$) is an indicator for Black children. We use the same specification as in Fryer and Levitt (2013b), excluding interviewer fixed effects. The control variables ($X$) include extensive information on sociodemographic characteristics, the home environment, and the prenatal environment (see their table 1B for descriptive statistics). After removing collinear terms there are $p = 78$ controls. VI. Implications for Inference and Comparison to High-Dimensional OLS-Based Methods The OVBs have important consequences for making inferences based on post–double Lasso.
the coverage rates of 90% confidence intervals based on
the DGPs in section VA and shows that the OVB of post-
double Lasso can cause substantial undercoverage even when
\( n = 1,000 \) and \( k = 5 \). The most important determinant of the
undercoverage is the sparsity parameter \( k \). Our results show
that the requirement on \( k \) for guaranteeing a good finite sam-
ple coverage accuracy for all \( R^2 \) can be quite stringent.

These results prompt the question of how to make infer-
ence in a reliable manner when one is concerned about OVBs.
In many economic applications, \( p \) is comparable to but still
smaller than \( n \). In such settings, OLS-based inference pro-
cedures provide a natural alternative to Lasso-based methods.
Under classical conditions, OLS is the best linear unbiased
estimator and admits exact finite sample inference as long
as \( p + 1 \leq n \), recalling that the number of regression co-
efficients is \( p + 1 \) in equation (1). Unlike the Lasso-based
inference methods, OLS does not rely on any sparsity as-
sumptions. This is important because sparsity assumptions
may not be satisfied in applications, and as we show in this
paper, the OVBs of Lasso-based inference procedures can be
substantial even when \( k \) is small and \( n \) is large and larger than
\( p \). Indeed, OLS-based inference exhibits desirable optimality
properties absent sparsity (or other restrictions) on \( \beta^* \).

While OLS is unbiased, constructing standard errors is
challenging when \( p \) is large. For instance, Cattaneo et al.
(2018) show that conventional Eicker-White robust standard
errors are inconsistent under asymptotics where \( p \) grows as
fast as \( n \). This result motivates a recent literature to develop
high-dimensional OLS-based inference procedures that are

| Method                        | Point estimate | Robust std. error |
|-------------------------------|----------------|-------------------|
| Post–double Lasso (\( \lambda_{BCCH} \)) | -0.6770        | 0.0114            |
| Post–double Lasso (0.5\( \lambda_{BCCH} \)) | -0.6762        | 0.0114            |
| Post–double Lasso (1.5\( \lambda_{BCCH} \)) | -0.6778        | 0.0114            |
| OLS with all controls         | -0.6694        | 0.0115            |
| OLS without controls          | -0.8538        | 0.0105            |

8For one-sided testing problems, the one-sided \( t \)-test based on OLS with
all controls is the uniformly most powerful test; for two-sided problems, the
two-sided \( t \)-test is the uniformly most powerful unbiased test (van der Vaart,
1998). We refer to section 4 in Armstrong and Kolesar (2016), section 5.5 in
Elliott, Müller, and Watson (2015), and section 2.1 in Li and Müller (2021)
for further discussions.
valid in settings with many controls (Cattaneo et al., 2018; Jochmans, 2022; Kline et al., 2020).

Figure 11 compares the finite sample performance of post–double Lasso and OLS with the heteroskedasticity robust HCK standard errors proposed by Cattaneo et al. (2018). Panel a shows that OLS exhibits close-to-exact empirical coverage rates irrespective of the magnitude of the coefficients and the implied $R^2$. The additional simulation evidence in appendix C confirms the excellent performance of OLS with HCK standard errors. Panel b displays the average length of 90% confidence intervals and shows that OLS yields somewhat wider confidence intervals than post–double Lasso.

In sum, our simulation results suggest that modern OLS-based inference methods that accommodate many controls may constitute a viable alternative to Lasso-based inference methods. These methods are unbiased and demonstrate an excellent size accuracy, irrespective of the magnitude of the coefficients corresponding to the relevant controls. However, there is a trade-off because OLS yields somewhat wider confidence intervals than post–double Lasso.

Finally, it is worth noting that we consider settings where one can easily invert $X^T X$ and the OLS and HCK variance estimators are numerically stable. In the case of singular or nearly singular $X^T X$, regularization is often unavoidable. Section VII provides a discussion of alternatives to OLS and Lasso-based inference methods.

VII. Recommendations for Empirical Practice

Here we summarize the practical implications of our results and provide guidance for empirical researchers.

First, the simulation evidence in section V and appendix C along with the theoretical results (see section IVC), suggest the following heuristic: if the estimates of $\alpha^*$ are robust to increasing the theoretically recommended regularization parameters in the two Lasso steps, post–double Lasso could be a reliable and efficient method. Therefore, we recommend to always check whether empirical results are robust to increasing the regularization parameters. Based on our simulations, a simple rule of thumb is to increase by 50% the regularization parameters proposed in Belloni et al. (2014). Robustness checks are standard in other contexts (e.g., bandwidth choices in regression discontinuity designs), and our results highlight the importance of such checks in the context of Lasso-based inference methods.
Second, following Belloni et al. (2014), we recommend to always augment the union of selected controls with an “amelioration” set of controls motivated by economic theory and prior knowledge to mitigate the OVBs.

Third, our simulations show that in moderately high-dimensional settings where \( p \) is comparable to but smaller than \( n \), recently developed OLS-based inference methods that are robust to the inclusion of many controls exhibit better size properties. These simulation results suggest that high-dimensional OLS-based procedures constitute a possible alternative to Lasso-based inference methods.

Fourth, OLS-based methods are not applicable when \( p > n \), and the OLS and variance estimators can be numerically unstable under severe multicollinearity even if \( p < n \). In such cases, regularization is often needed. Ridge regressions, which impose restrictions on the Euclidean norm of \( \beta^* \), avoid variable selection and may be a useful alternative to the Lasso. (See also Armstrong et al., 2020, for related restrictions on \( \beta^* \).)

Finally, in many economic applications, researchers start with a small number of raw controls and want to use a flexible nonparametric model to capture the dependence of outcomes on controls while maintaining a simple parametric form for modeling the variables of interest. Such a specification leads to the classical partially linear models. In fact, Belloni et al. (2014) motivate post–double Lasso with these models. If one is concerned about OVBs, inference methods that do not rely on variable selection are natural alternatives to post–double Lasso. Under suitable smoothness restrictions on the nonparametric component, inference on the parameter of interest in partially linear models is a well-studied problem (Robinson, 1988; Newey & McFadden, 1994). The frameworks proposed in these papers can be built on procedures such as sieves (Chen, 2007), local nonparametric methods (Fan & Gijbels, 1996), and kernel ridge regressions (Schölkopf & Smola, 2002).

VIII. Conclusion

Given the rapidly increasing popularity of Lasso and Lasso-based inference methods in empirical economic research, it is crucial to better understand the merits and
limitations of these new tools and how they compare to other alternatives such as the high-dimensional OLS-based procedures.

This paper presents theoretical results as well as simulation and empirical evidence on the finite sample behavior of post–double Lasso and the debiased Lasso (in the appendix). Specifically, we analyze the finite sample OVBs arising from the Lasso not selecting all the relevant control variables. Our results have important practical implications, and we provide guidance for empirical researchers.

We focus on the implications of underselection for post–double Lasso and the debiased Lasso in linear regression models. However, our results on the underselection of the Lasso also have important implications for other inference methods that rely on Lasso as a first-step estimator. Toward this end, an interesting avenue for future research would be to investigate the impact of underselection on the performance of the Lasso-based approaches proposed by Belloni et al. (2014), Farrell (2015), Belloni et al. (2017b), and Chernozhukov et al. (2018) for nonlinear models. In moderately high-dimensional settings where $p$ is smaller than but comparable to $n$, it would also be interesting to compare the treatment effects estimators in Belloni et al. (2014) to the robust finite sample methods proposed by Rothe (2017).

Finally, this paper motivates further examinations of the practical usefulness of Lasso-based inference procedures and other modern high-dimensional methods. For example, Angrist and Frandsen (2019) present interesting simulation evidence on the finite sample behavior of Lasso-based IV methods (Belloni et al., 2012). It would be interesting to explore the implications of our theoretical results on the underselection of the Lasso in problems with weak instruments.

REFERENCES

Abadie, Alberto, “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” Journal of Econometrics 113:2 (2003), 231–263. 10.1016/S0304-4076(02)00201-4

Angrist, Joshua D., and Brigham Frandsen, “Machine Labor,” NBER working paper 26584 (2019).

Armstrong Timothy, and Michal Kolesar, Optimal Inference in a Class of Regression Models, arXiv:1511.06028v2 (2016).

Armstrong, Timothy B., Michal Kolesár, and Soonwoo Kwon, Bias-Aware Inference in Regularized Regression Models, arXiv:2012.14823 (2020).

Belloni, Alexandre, and Victor Chernozhukov, “Least Squares after Model Selection in High-Dimensional Sparse Models,” Bernoulli 19:2 (2013), 521–547. 10.3150/11-BEJ410

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain,” Econometrica 80:6 (2012), 2369–2429. 10.3982/ECTA9626

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, “Inference on Treatment Effects after Selection among High-Dimensional Controls,” Review of Economic Studies 81:2 (2014), 608–650. 10.1093/restud/rdt044

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen, “Supplement to ’Program Evaluation and Causal Inference with High-Dimensional Data‘,” Econometrica Supplemental Materials (2017a).

“Program Evaluation and Causal Inference with High-Dimensional Data,” Econometrica 85:1 (2017b), 233–298.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur, “Inference in High-Dimensional Panel Models with an Application to Gun Control,” Journal of Business and Economic Statistics 34:4 (2016), 590–605. 10.1080/07350015.2015.1102733

Benjamin, Daniel J., “Does 401(k) Eligibility Increase Saving? Evidence from Propensity Score Subclassification,” Journal of Public Economics 87:5 (2003), 1259–1290. 10.1016/S0047-2727(01)00167-0

Bickel, Peter J., Ya’acov Ritov, and Alexandre B. Tsybakov, “Simultaneous Analysis of Lasso and Dantzig Selector,” Ann. Statist. 37:4 (2009), 1705–1732. 10.1214/08-AOS626

Breza, Emily, and Arun G. Chandrasekhar, “Social Networks, Reputation, and Commitment: Evidence from a Savings Monitors Experiment,” Econometrica 87:1 (2019), 175–216. 10.3982/ECTA13683

Caner, Mehmet, and Anders Bredahl Kock, “Asymptotically Honest Confidence Regions for High Dimensional Parameters by the Desparsified Conservative Lasso,” Journal of Econometrics 203:1 (2018), 143–168. 10.1016/j.jeconom.2017.11.005

Cattaneo, Mattias D., Michael Jansson, and Whitney K. Newey, “Inference in Linear Regression Models with Many Covariates and Heteroscedasticity,” Journal of the American Statistical Association 113:523 (2018), 1350–1363. 10.1080/01621459.2017.1329360

Chen, Daniel L., “Can Markets Stimulate Rights? On the Alienability of Legal Claims,” RAND Journal of Economics 46:1 (2015), 23–65. 10.1111/1756-2171.12076

Chen, Xiaohong, “Large Sample Sieve Estimation of Semi-Nonparametric Models” (pp. 5549–5632), in James Heckman and Edward E. Leamer, eds., Handbook of Econometrics, vol. 6 (Amsterdam: Elsevier, 2007).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Dufo, Christian Hansen, Whitney Newey, and James Robins, “Double/Debiased Machine Learning for Treatment and Structural Parameters,” Econometrics Journal 21:1 (2018), 1–68. 10.1111/eca.12097

Chernozhukov, Victor, Ivan Fernandez-Val, Blaise Melly, and Kaspar Wüthrich, “Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes,” Journal of the American Statistical Association 115:529 (2020), 123–137. 10.1080/01621459.2019.1611581

Chernozhukov, Victor, Christian Hansen, and Martin Spindler, “HDM: High-Dimensional Metrics,” R Journal 8:2 (2016), 185–199. 10.32614/RJ-2016-040

Chernozhukov, Victor, Christian Hansen, and Christian Hansen, “The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis,” this review 86:3 (2004), 735–751.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov, “On Cross-Validated Lasso in High Dimensions,” Annal Stat. 40:3 (2021), 1300–1317.

Cole, Shawn, and Nilesh A. Fernando, “Mobile’izing Agricultural Advice Technology Adoption Diffusion and Sustainability,” Economic Journal 131:633 (2020), 192–219. 10.1093/ej/ueaa084

D’Adamo, Riccardo, Cluster-Robust Standard Errors for Linear Regression Models with Many Controls, arXiv:1806.07314 (2018).

Decker, Simon, and Hendrik Schmitz, “Health Shocks and Risk Aversion,” Journal of Health Economics 50 (2016), 156–170. 10.1016/j.jhealeco.2016.09.006. PubMed: 27792902

Elliott, Graham, Ulrik K. Møller, and Mark W. Watson, “Nearly Optimal Tests When a Nuisance Parameter Is Present under the Null Hypothesis,” Econometrica 83:2 (2015), 771–711. 10.3982/ECTA10535

Enke, Benjamin, “Moral Values and Voting,” Journal of Political Economy 128:10 (2020), 3679–3729. 10.1086/708857

Fan, Jianqing, and Irene Gijbels, Local Polynomial Modelling and Its Applications (Boca Raton, FL: Routledge, 1996).

Farrell, Max H., “Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations,” Journal of Econometrics 189:1 (2015), 1–23. 10.1016/j.jeconom.2015.06.017

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent,” Journal of Statistical Software 33:1 (2010), 1–22. 10.18637/jss.v033.i01. PubMed: 20808728

Fryer, Roland G., and Steven D. Levitt, Replication Data for: Testing for Racial Differences in the Mental Ability of Young Children (Nashville, TN: American Economic Association, 2013a). Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 2019-10-11. 10.3886/E112609V1

——— “Program Evaluation and Causal Inference with High-Dimensional Data,” Econometrica 85:1 (2017b), 233–298.
