cemmap

centre for microdata methods and practice

Estimation of treatment effects with high-dimensional controls

A. Belloni
V. Chernozhukov
C. Hansen

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

ESTIMATION OF TREATMENT EFFECTS WITH HIGH-DIMENSIONAL CONTROLS

A. BELLONI, V. CHERNOZHUKOV, AND C. HANSEN

ABSTRACT. We propose methods for inference on the average effect of a treatment on a scalar outcome in the presence of very many controls. Our setting is a partially linear regression model containing the treatment/policy variable and a large number p of controls or series terms, with p that is possibly much larger than the sample size n, but where only $s \ll n$ unknown controls or series terms are needed to approximate the regression function accurately. The latter sparsity condition makes it possible to estimate the entire regression function as well as the average treatment effect by selecting an approximately the right set of controls using Lasso and related methods. We develop estimation and inference methods for the average treatment effect in this setting, proposing a novel "post double selection" method that provides attractive inferential and estimation properties. In our analysis, in order to cover realistic applications, we expressly allow for imperfect selection of the controls and account for the impact of selection errors on estimation and inference. In order to cover typical applications in economics, we employ the selection methods designed to deal with non-Gaussian and heteroscedastic disturbances. We illustrate the use of new methods with numerical simulations and an application to the effect of abortion on crime rates.

Key Words: treatment effects, high-dimensional regression, inference under imperfect model selection

1. INTRODUCTION

Many empirical analyses in economics focus on estimating the structural, causal, or treatment effect of some variable on an outcome of interest. For example, we might be interested in the estimating the causal effect of the minimum wage or some other government policy on employment. Since economic policies and many other economic variables are not randomly assigned, economists rely on a variety of quasi-experimental approaches based on observational data when trying to estimate such effects. One popular method is based on the assumption that the variable of interest can be taken as randomly assigned once a sufficient set of other

factors has been controlled for. Economists, for example, might argue that deviations in state-level minimum wages can be taken as randomly assigned relative to unobservable factors that could affect state-level employment once aggregate macroeconomic activity, state-level economic activity, and state-level demographics have been controlled for; see Card and Krueger (1997) among other references.

A problem empirical researchers face when relying on an identification strategy for estimating a structural effect that relies on a conditional on observables argument is knowing which variables to control for. Typically, economic intuition will suggest a set of variables that might be important but will not identify exactly which variables are important or the functional form with which variables should enter the model. This lack of clear guidance about what variables to use leaves researchers with the problem of attempting to select a sensible set of controls from a potentially vast set of control variables including raw regressors available in the data as well as interactions and other transformations of these regressors. A typical economic study will rely on a sensitivity analysis in which a researcher reports results for several different sets of controls in an attempt to show that the parameter of interest that summarizes the causal effect of the policy variable is insensitive to changes in the set of control variables. See Donohue III and Levitt (2001), which we use as the basis for the empirical study in this paper, or examples in Angrist and Pischke (2008) among many other references.

In this paper, we present an approach to estimating structural effects in an environment where we believe that the treatment variable may be taken as exogenous conditional on observables that complements existing strategies. We pose the problem in the framework of a partially linear model

$$y_{1i} = d_i \alpha_0 + g(z_i) + \zeta_i$$

where d_i is the treatment variable of interest, z_i is a set of control variables, and ζ_i is an unobservable that satisfies $\mathrm{E}[\zeta_i | d_i, z_i] = 0$. This model is general enough to accommodate the usual models used in estimating treatment effects in applied economic research. Within this model, the problem we examine is selecting a small set of variables from among z_i and potentially transformations of z_i to adequately approximate $g(z_i)$ and make estimation and inference about the parameter of interest α_0 feasible. We allow for selection among a large set of p observable variables consisting of z_i and transformations where $p \gg n$ is allowed. This framework allows for the realistic scenario in which the researcher is unsure about exactly which variables or transformations are important for approximating $g(z_i)$ and so is left with searching among a broad set of controls.

Of course, without further structure on the data, useful inference about α_0 is unavailable. We impose such structure by assuming that among the very large set of potential conditioning variables, there is a relatively small set consisting of $s < n$ variables whose identities are *a priori* unknown by the researcher that provide a good enough approximation that the exogeneity of d_i may be taken as given once these variables have been controlled for. This assumption, which is termed sparsity, allows us to approach the estimation problem as a variable selection problem from among a large set of controls.

Sparsity corresponds quite well to usual approaches to conditional on observable analyses in applied economics where the set of sensitivity analyses reported generally rely on estimating the treatment effect considering different small sets of potential control variables, sets with far fewer variables than there are observations in the sample. We consider a formal approach to variable selection in this setting that complements the usual *ad hoc* approaches based on variable selection using ℓ_1-penalization methods, especially Lasso.

ℓ_1-penalized methods have been proposed for model selection problems in high-dimensional least squares problems (Tibshirani 1996) in part because they are computationally efficient (avoiding a curse of dimensionality). Recently, many of these methods have been shown to have good estimation properties even when perfect variable selection is not feasible (e.g. Candes and Tao (2007), Meinshausen and Yu (2009), Bickel, Ritov, and Tsybakov (2009), Belloni and Chernozhukov (2011c) and the references therein). Such methods were also shown to extend suitably to nonparametric and non-Gaussian cases (e.g. Bickel, Ritov, and Tsybakov (2009), Belloni, Chen, Chernozhukov, and Hansen (2010)). Also, these methods produce models with a relatively small set of variables. The last property is important in that it leaves the researcher with a set of variables that may be examined further in addition to corresponding to the usual approach in economics that relies on considering a relatively small number of controls.

A main contribution of this paper is providing theory that gives conditions under which ℓ_1-penalized estimators may be successively used to estimate structural economic effects of interest and in offering a simple and robust method to estimating these effects. The approach we advocate differs from usual uses of Lasso-type methods by relying on two different variable selection steps. In the first, we select a set of control variables that are useful for predicting the treatment d_i. This step helps to insure robustness by finding control variables that are strongly related to the treatment and thus potentially important compounds. We then select additional variables by selecting control variables that predict y_{1i}. This step helps to insure that we have captured important elements in the equation of interest, ideally helping keep the residual variance small as well as intuitively providing an additional chance to find important

confounds. The treatment effect of interest is then estimated by the linear regression of y_{1i} on the treatment d_i and the union of the set of variables selected in the two previous steps. We provide theoretical results on the properties of the resulting treatment effect estimator and show that it may achieve the semi-parametric efficiency bound under some conditions. Importantly, our theoretical results allow for imperfect variable selection in either of the two variable selection steps as well as allowing for non-Gaussianity and heteroskedasticity of the model's errors. [1]

We illustrate the theoretical results through an examination of the effect of abortion on crime rates following Donohue III and Levitt (2001). In this example, we find that the formal variable selection procedure produces a qualitatively different result than that obtained through the *ad hoc* set of sensitivity results presented in the paper. By using formal variable selection, we select a small set of between eight and fourteen variables depending on the outcome, as opposed to the set of six variables considered by Donohue III and Levitt (2001). Once this set of variables is linearly controlled for, the estimated abortion effect is rendered extremely imprecise. It is interesting that the key variable selected by the variable selection procedure is the initial condition for the abortion rate. This selection and the resulting imprecision of the estimated treatment effect suggests that one cannot determine precisely whether the effect attributed to abortion without including this initial condition is due to changes in the abortion rate or some other persistent state-level factor that is related to relevant changes in the abortion rate and current changes in the crime rate.[2] Finding that a simple-to-implement, formal approach to variable selection produces a qualitatively different result than a more *ad hoc* approach suggests that there is room for such procedures in applied economics and that these methods might be used to complement economic intuition in selecting control variables for estimating treatment effects in settings where treatment is taken as exogenous conditional on observables.

Notation. In what follows, we work with triangular array data $\{(\omega_{i,n}, i = 1, ..., n), n = 1, 2, 3, ...\}$ defined on some common probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Each $\omega_{i,n} = (y'_{i,n}, z'_{i,n}, d'_{i,n})'$ is a vector, with components defined below in what follows, and these vectors are i.n.i.d. – independent across i, but not necessarily identically distributed. The law of $\{\omega_{i,n}, i = 1, ..., n\}$ can change with n. Thus, all parameters that characterize the distribution of $\{\omega_{i,n}, i = 1, ..., n\}$ are implicitly indexed by the sample size n, but we omit the explicit index n in

[1]In a companion work (Belloni, Chernozhukov, and Hansen 2011) we have obtained similar results in the ideal Gaussian homoscedastic framework.

[2]Note that all models are estimated in first-differences to eliminate any state-specific factors that might be related to both the relevant level of the abortion rate and the level of the crime rate.

what follows to simplify notation. We use array asymptotics to better capture some finite-sample phenomena and to retain the robustness of conclusions to perturbations of the data-generating process. We also use the following empirical process notation, $\mathbb{E}_n[f] := \mathbb{E}_n[f(\omega_i)] := \sum_{i=1}^n f(\omega_i)/n$, and $\mathbb{G}_n(f) := \sum_{i=1}^n (f(\omega_i) - \mathrm{E}[f(\omega_i)])/\sqrt{n}$. Since we want to deal with i.n.i.d. data, we also introduce the average expectation operator: $\bar{\mathrm{E}}[f] := \mathrm{E}\mathbb{E}_n[f] = \mathrm{E}\mathbb{E}_n[f(\omega_i)] = \sum_{i=1}^n \mathrm{E}[f(\omega_i)]/n$, The l_2-norm is denoted by $\|\cdot\|$, and the l_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use $\|\cdot\|_\infty$ to denote the maximal element of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1,\ldots,p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We use the notation $(a)_+ = \max\{a,0\}$, $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E, we say that E wp $\to 1$ when E occurs with probability approaching one as n grows. Given a p-vector b, we denote $\mathrm{support}(b) = \{j \in \{1,...,p\} : b_j \neq 0\}$.

2. Inference on Treatment and Structural Effects Conditional on Observables

2.1. Framework.
In this paper we consider the following partially linear model

$$(2.1) \qquad y_{1i} = d_i\alpha_0 + g(z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0,$$

$$(2.2) \qquad d_i = m(z_i) + v_i, \qquad \mathrm{E}[v_i \mid z_i] = 0,$$

where y_{1i} is the outcome variable, d_i is the policy/treatment variable whose impact α_0 we would like to infer, z_i represents confounding factors on which we need to condition, and ζ_i and v_i are disturbances. Under appropriate conditions, the parameter α_0 is the average treatment or structural effect, Heckman, LaLonde, and Smith (1999) and Imbens (2004), and is of major interest in many empirical studies.

The confounding factors z_i affect the policy variable via the function $m(z_i)$ and the outcome variable via function $g(z_i)$. Both of these functions are unknown and potentially complicated. We use linear combinations of (possibly technical) control terms $x_i = P(z_i)$ to approximate $g(z_i)$ and $m(z_i)$, writing (2.1) and (2.2) as

$$(2.3) \qquad y_{1i} = d_i\alpha_0 + \underbrace{x_i'\beta_{g0} + r_{gi}}_{g(z_i)} + \zeta_i,$$

$$(2.4) \qquad d_i = \underbrace{x_i'\beta_{m0} + r_{mi}}_{m(z_i)} + v_i,$$

where $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ are some approximations to $g(z_i)$ and $m(z_i)$, and r_{gi} and r_{mi} are the corresponding approximation errors. In order to allow for a flexible specification and incorporation of all pertinent confounding factors, the vector of controls, $x_i = P(z_i)$, can have a dimension $p = p_n$ which can be high in relation to the sample size. In fact, p can be possibly much larger than the sample size n though restricted via $\log p = o(n^{1/3})$ and via other conditions stated below. For example, high-dimensional instruments $x_i = P(z_i)$ could arise as any combination of the following two cases:

- **Many controls.** The list of available controls is large, in which case we have $x_i = z_i$, as in e.g. Koenker (1988).
- **Many technical controls.** The list $x_i = P(z_i)$ consists of a large number of series terms with respect to some elementary regressor vector z_i, e.g., x_i could be composed of B-splines, dummies, polynomials, and various interactions as in e.g. (Newey 1997).

The high-dimensional p creates a challenge, which is particularly apparent when $p > n$. However, a key condition that makes it possible to perform constructive estimation and inference in such cases is sparsity, namely that there exist sparse approximations $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ to $g(z_i)$ and $m(z_i)$ in (2.3)-(2.4) that render the approximation errors r_{gi} and r_{mi} sufficiently small. Or, more formally, there exist β_{g0} and β_{m0} such that at most $s = s_n \ll n$ elements of β_{m0} and β_{g0} are non-zero, namely

$$\|\beta_{m0}\|_0 \leqslant s \text{ and } \|\beta_{g0}\|_0 \leqslant s,$$

where identities of these elements are unknown, and where the size of the resulting approximation errors is small compared to the conjectured size of the estimation error:

$$\{\mathbb{E}_n[r_{gi}^2]\}^{1/2} \lesssim_P \sqrt{s/n} \text{ and } \{\mathbb{E}_n[r_{mi}^2]\}^{1/2} \lesssim_P \sqrt{s/n}.$$

In other words, out of potentially many controls x_i only at most $s = s_n \ll n$ unknown controls are sufficient for approximating the functions $m(z_i)$ and $g(z_i)$ well enough. Note that the size of the approximating model $s = s_n$ can grow with n, just as in the standard series estimation or estimation with many regressors.

The frameworks above extends the standard framework in the treatment effect literature which assumes both that the relevant controls are known and that the number of such controls s is much smaller than the sample size. Here we instead assume that there are many, p, potential controls of which at most s controls are important, and the identity of these controls is unknown. Relying on such sparsity assumption, we shall employ model selection methods to

select at least approximately the right set of controls and then estimate the treatment effect α_0.

2.2. **The Method: Least Squares after Double Selection.** We propose the following method for estimation and inference on α. The most important and novel feature of this method is that it does not rely on the highly unrealistic assumption of perfect model selection, which is often invoked to justify inference after model selection. [3] Moreover, its (non-apparent) construction reflects our effort to offer a method that has attractive robustness features, providing estimator that is $\sqrt{n}$ consistent and asymptotically normal under mild conditions, and providing confidence intervals that are robust to various perturbations of the data-generating process that preserve approximate sparsity.

To define the method, we first write the reduced form corresponding to (2.1)-(2.2) as:

$$(2.5) \qquad\qquad y_{1i} = x_i'\bar{\beta}_0 + \bar{r}_i + \bar{\zeta}_i,$$

$$(2.6) \qquad\qquad d_i = x_i'\beta_{m0} + r_{mi} + v_i,$$

where $\bar{\beta}_0 := \alpha_0\beta_{m0} + \beta_{g0}, \;\; \bar{r}_i := \alpha_0 r_{mi} + r_{gi}, \;\; \bar{\zeta}_i := \alpha_0 v_i + \zeta_i.$

Now we have two equations and hence can apply model selection methods to each equation to select control terms. The chief method we will use will be the Lasso method described in more detail below. Then we can run least squares of y_{1i} on d_i and the union of the controls selected in each equation to estimate and perform standard inference on α_0. Intuitively, we are more likely to recover key controls by considering selection of controls from both equations instead of just considering selection of controls from a single equation such as (2.1), (2.3), or (2.4). In the various finite-sample experiments, we show that none of such "single selection" methods work as well as the double selection method. Theoretically this is also supported by the fact that the double selection method requires much weaker regularity conditions for its validity and for attaining the efficiency bound[4] than single selection methods.

Now we formally define the post double selection estimator: Let $\widehat{I}_1 = \text{support}(\widehat{\beta}_1)$ denote the control terms selected by a feasible Lasso estimator $\widehat{\beta}_1$ computed using data $(\tilde{y}_i, \tilde{x}_i) = (d_i, x_i), i = 1, ..., n$. Let $\widehat{I}_2 = \text{support}(\widehat{\beta}_2)$ denote the control terms selected by a feasible Lasso

---

[3]To the best of our knowledge this result is a first result of this kind, as it pertains to our setting. This result extends our previous results on inference under imperfect model selection in the instrumental regression model (Belloni, Chernozhukov, and Hansen 2010, Belloni, Chen, Chernozhukov, and Hansen 2010) and in partially linear Gaussian model (Belloni, Chernozhukov, and Hansen 2011). It should ne noted that the analysis here is considerably more involved.

[4]Semi-parametric efficiency is attained in the homoscedastic cases.

estimator $\widehat{\beta}_2$ computed using data $(\tilde{y}_i, \tilde{x}_i) = (y_{1i}, x_i), i = 1, ..., n$. Finally, the post double selection estimator $\check{\alpha}$ of α_0 is defined as the least squares estimator obtained by regressing y_{1i} on d_i and the selected control terms x_{ij} with $j \in \widehat{I} \supseteq \widehat{I}_1 \cup \widehat{I}_2$:

$$(\check{\alpha}, \check{\beta}) = \operatorname*{argmin}_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \{\mathbb{E}_n[(y_{1i} - d_i\alpha - x_i'\beta)^2] \; : \; \beta_j = 0, \forall j \notin \widehat{I}\}.$$

The set $\widehat{I}$ may in addition contain other variables with names $\widehat{I}_3$ that the analyst may think are important for ensuring robustness. We call $\widehat{I}_3$ the amelioration set. Thus, $\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2 \cup \widehat{I}_3$; let $\widehat{s} = |\widehat{I}|$ and $\widehat{s}_j = |\widehat{I}_j|$ for $j = 1, 2, 3$.

We define feasible Lasso estimator below and note that other selection methods can be used as well under conditions specified in Section 5. When the feasible Lasso is used we shall refer to the post double selection estimator as the post double Lasso estimator.

The main theoretical result of the paper shows that the post-double-selection estimator $\check{\alpha}$ obeys

$$([\bar{\mathbb{E}}v_i^2]^{-1}\bar{\mathbb{E}}[v_i^2\zeta_i^2][\bar{\mathbb{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) \to_d N(0, 1)$$

under approximate sparsity conditions, and uniformly in a rich set of data-generating processes. Moreover, we provide the consistent standard errors based on the plug-in principle.

2.3. **Selection of controls via feasible Lasso Methods.** Here we describe feasible selection via Lasso. Note that each of the regression equations above is of the form

$$\tilde{y}_i = \underbrace{\tilde{x}_i'\beta_0 + r_i}_{f(\tilde{z}_i)} + \epsilon_i,$$

where $f(\tilde{z}_i)$ is the regression function, $\tilde{x}_i'\beta_0$ is the approximation based on the dictionary $\tilde{x}_i = P(\tilde{z}_i)$, r_i is the approximation error, and ϵ_i is the error. Tibshirani (1996) propose the Lasso estimator/model selector defined as a solution to

$$(2.7) \qquad\qquad \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2] + \frac{\lambda}{n}\|\beta\|_1,$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. The kinked nature of the penalty function forces the solution $\widehat{\beta}$ to have many zeroes, which has convenient model selection applications. The selected model $\widehat{T} = \text{support}(\widehat{\beta})$ is often used for further refitting by least squares, leading to the so called post-Lasso or Gauss-Lasso estimator, see, e.g., Belloni and Chernozhukov (2011c). The Lasso estimator/selector is computationally attractive because it minimizes a convex function. In the homoskedastic case, a basic choice for penalty level suggested by Bickel, Ritov, and Tsybakov

(2009) is

$$\lambda = 2 \cdot c\sigma\sqrt{2n\log(2p/\gamma)}, \tag{2.8}$$

where $c > 1$ and $1-\gamma$ is a confidence level that needs to be set close to 1. The formal motivation for this penalty is that it leads to near-oracle rates of convergence of the estimator under approximate sparsity. This in turn implies good approximation properties of the selected model $\widehat{T}$, as noted in Belloni and Chernozhukov (2011c). Unfortunately, even in the homoskedastic case the penalty level specified above is not feasible since it depends on the unknown σ.

Belloni, Chen, Chernozhukov, and Hansen (2010) formulate a feasible Lasso estimator/selector $\widehat{\beta}$ geared for heteroscedastic, non-Gaussian cases, which solves

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2] + \frac{\lambda}{n}\|\widehat{\Psi}\beta\|_1, \tag{2.9}$$

where $\widehat{\Psi} = \mathrm{diag}(\widehat{l}_1, \ldots, \widehat{l}_p)$ is a diagonal matrix specifying penalty loadings. The penalty level λ and loadings $\widehat{l}_j$'s are set

$$\lambda = 2 \cdot c\Phi^{-1}(1-\gamma/2p) \text{ and } \widehat{l}_j = l_j + o_P(1), \ l_j = \sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2\epsilon_i^2]}, \text{ uniformly in } j = 1, \ldots, p, \tag{2.10}$$

where $c > 1$ and $1 - \gamma$ is a confidence level.[5] Since "ideal" l_j's are not observed, they are estimated by $\widehat{l}_j$ via an iteration method defined in Appendix A. We refer to the resulting feasible Lasso method as the *Iterated Lasso*. The estimator $\widehat{\beta}$ has statistical performance that is similar to that of the (infeasible) Lasso described above in the Gaussian cases, but also delivers Gaussian-like performance in the non-Gaussian, heteroscedastic case (Belloni, Chen, Chernozhukov, and Hansen 2010). In our case, we only use $\widehat{\beta}$ for purposes of model selection, namely we use

$$\widehat{T} = \mathrm{suppport}(\widehat{\beta}),$$

the labels of regressors for which estimated coefficients are zero. The selected model has good approximation properties under approximate sparsity, as we formally state below in Section 3.

Belloni, Chernozhukov, and Wang (2011) propose another feasible variant of Lasso called the *Square-root Lasso* estimator $\widehat{\beta}$ defined as a solution to

$$\min_{\beta \in \mathbb{R}^p} \sqrt{\mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2]} + \frac{\lambda}{n}\|\widehat{\Psi}\beta\|_1, \tag{2.11}$$

with the penalty level

$$\lambda = c \cdot \Phi^{-1}(1 - \gamma/2p). \tag{2.12}$$

---

[5]Practical recommendations include the choice $c = 1.1$ and $\gamma = .05$.

where $c > 1$ and $\gamma \in (0,1)$ is the confidence level, and $\widehat{\Psi}$ is a diagonal matrix containing penalty loadings $\widehat{l}_1, ..., \widehat{l}_p$ in the diagonal. The main attractive feature of (2.11) is that in the homoscedastic case we can set $\widehat{l}_j = \{\mathbb{E}_n[\tilde{x}_{ij}^2]\}^{1/2}$, and the penalty level λ is independent of the value $\mathrm{E}[\epsilon_i^2] = \sigma^2$, and so it is pivotal. In the heteroscedastic case, we would like to choose

$$(2.13) \quad l_j + o_P(1) \leqslant \widehat{l}_j \lesssim_P l_j, \text{ where } l_j = \{\mathbb{E}_n[\tilde{x}_{ij}^2 \epsilon_i^2]]/\mathbb{E}_n[\epsilon_i^2]\}^{1/2}, \text{ uniformly in } j = 1, ..., p.$$

For example, since $\{\mathbb{E}_n[\tilde{x}_{ij}^2 \epsilon_i^2]]/\mathbb{E}_n[\epsilon_i^2]\}^{1/2} \leqslant \{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/4}\{\mathbb{E}_n[\epsilon_i^4]\}^{1/4}/\{\mathbb{E}_n[\epsilon_i^2]\}^{1/2}$, we can use $\widehat{l}_j = \{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/4}2$, which gives $l_j + o_P(1) \leqslant \widehat{l}_j$ if $\{\mathbb{E}_n[\epsilon_i^4]\}^{1/4}/\{\mathbb{E}_n[\epsilon_i^2]\}^{1/2} \leqslant 2 + o_P(1)$, which covers a wide class of marginal distributions for error ϵ_i, for example, all t-distributions with degree of freedom greater than five. As in the previous case, we can iteratively re-estimate the penalty loadings to obtain the refined penalty loadings:

$$(2.14) \qquad\qquad \widehat{l}_j = l_j + o_P(1), \text{ uniformly in } j = 1, ..., p.$$

The resulting Lasso and post-Lasso estimators based on this have attractive Gaussian-like performance even in non-Gaussian, heteroscedastic cases. This implies good approximation properties for the selected model $\widehat{T}$.

In what follows, the name *feasible Lasso* will be formally used to name either the Iterated Lasso estimator $\widehat{\beta}$ solving (2.9)-(2.10) or Square-root Lasso estimator $\widehat{\beta}$ solving (2.11)-(2.13), with the confidence level $1 - \gamma$ such that

$$(2.15) \qquad\qquad \gamma = o(1) \text{ and } \log(1/\gamma) \lesssim \log(p \vee n).$$

3. Theory of Estimation and Inference

3.1. Regularity Conditions.
In this section we record regularity conditions that are sufficient for validity of the main estimation and inference result. We begin by stating our main condition, which contains the previously defined approximate sparsity as well as other more technical assumptions.

Condition ASTE. *(i) For each n, the data array $D_n = \{(y_{1i}, d_i, z_i), i = 1, ..., n\}$ is a sequence of i.n.i.d vectors that obey the model (2.1)-(2.2) for each n, and $x_i = P(z_i)$ is a dictionary of transformations of z_i. We allow the law of data to change with n and so all parameters can depend on n. (ii) The parameter value α_0 is bounded uniformly in n. (iii) Functions m and g admit an approximately sparse form, with sparsity index s, namely there*

exists $s \geqslant 1$ and β_{m0} and β_{g0} such that

$$(3.16) \qquad m(z_i) = x_i' \beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{mi}^2]\}^2 \lesssim_P \bar{\sigma}\sqrt{s/n},$$

$$(3.17) \qquad g(z_i) = x_i' \beta_{g0} + r_{gi}, \qquad \|\beta_{g0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{gi}^2]\}^2 \lesssim_P \bar{\sigma}\sqrt{s/n},$$

where the parameter values are indexed by n. (iv) The sparsity index obeys $s^2 \log^2(p \vee n) = o(n)$ and the size of the amelioration set obeys $\widehat{s}_3 \lesssim 1 \vee \widehat{s}_1 \vee \widehat{s}_2$; and (v) $\max_{i \leqslant n}(d_i^2/s^2 + \|x_i\|_\infty^2)(|v_i|^2 + |\zeta_i|^2 + |r_{gi}|^2 + |r_{mi}|^2)s^2 \log(p \vee n) = o_P(n)$, and $\mathbb{E}_n[(v_i + r_{mi})^2(\zeta_i + r_{gi})^2] - \bar{\mathrm{E}}[v_i^2\zeta_i^2] \to_P 0$.

Comment 3.1. The condition (ASTE(i)) states formally the modeling assumption and imposes independent sampling on the data. For each n, the data vectors D_n are defined on some common probability space $(\Omega, \mathcal{F}, P)$. Even though there is a common underlying probability space for all n, we allow the law P_n of data array D_n to depend on n. In other words, we allow for triangular array sequences, which allows us to insure robustness to perturbations of the data generating process P_n. The approximate sparsity (ASTE(iii)) and the growth condition (ASTE(iv)) are the main conditions for establishing our main inferential result. Condition ASTE(iv) requires that the size $\widehat{s}_3$ of the amelioration set $\widehat{I}_3$ should be no larger than the size selected by the Lasso method. Simply put, if we decide to include controls in addition to those selected by Lasso, the total number of additions should not exceed (much more) than what was selected by Lasso. This will ensure that the total number $\widehat{s}$ of controls selected will obey $\widehat{s} \lesssim_P s$, and we require that $s^2 \log^2 p/n \to 0$. Condition ASTE(v) is simply a set of sufficient conditions for the consistent estimation of the variance of the double selection estimator (for instance, it is implied by the other conditions if regressors are uniformly bounded and the approximation errors are going to zero a.s.). $\qquad \square$

The next condition concerns the behavior of the Gram matrix $\mathbb{E}_n[x_i x_i']$. Whenever $p > n$, the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ does not have full rank and in principle is not well-behaved. However, we only need good behavior of smaller submatrices. Define the minimal and maximal m-sparse eigenvalue of a semi-definite matrix M as

$$(3.18) \qquad \phi_{\min}(m)[M] := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2} \quad \text{and} \quad \phi_{\max}(m)[M] := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}.$$

To assume that $\phi_{\min}(m)[\mathbb{E}_n[x_i x_i']] > 0$ requires that all empirical Gram submatrices formed by any m components of x_i are positive definite. We shall employ the following condition as a sufficient condition for our results.

Condition SE. *There is $\ell_n \to \infty$ such that the maxiamal and minimal $\ell_n s$-sparse eigenvalues are bounded from below and away from zero, namely*

$$\kappa' \leqslant \phi_{\min}(\ell_n s)[\mathbb{E}_n[x_i x_i']] \leqslant \phi_{\max}(\ell_n s)[\mathbb{E}_n[x_i x_i']] \leqslant \kappa'',$$

where $0 < \kappa' < \kappa'' < \infty$ are constants that do not depend on n.

Comment 3.2. It is well-known that Condition SE is quite plausible for many designs of interest. For instance, Condition SE holds with probability approaching one as $n \to \infty$ if x_i is a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij}/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$, and

- $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. zero-mean Gaussian random vectors that have population Gram matrix $\mathrm{E}[\tilde{x}_i\tilde{x}_i']$ with ones on the diagonal and its minimal and maximal $s \log n$-sparse eigenvalues bounded away from zero and from above, where $s \log n = o(n/\log p)$;
- $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. bounded zero-mean random vectors with $\|\tilde{x}_i\|_\infty \leqslant K_n$ a.s. that have population Gram matrix $\mathrm{E}[\tilde{x}_i\tilde{x}_i']$ with ones on the diagonal and its minimal and maximal $s \log n$-sparse eigenvalues bounded from above and away from zero, where $K_n^2 s \log^5(p \vee n) = o(n)$.

Recall that a standard assumption in econometric research is to assume that the population Gram matrix $\mathrm{E}[x_i x_i']$ has eigenvalues bounded from above and below, see e.g. Newey (1997). The conditions above allow for this and more general behavior, requiring only that the $s \log n$ sparse eigenvalues of the population Gram matrix $\mathrm{E}[x_i x_i']$ are bounded from below and from above. The latter is important for allowing functions x_i to be formed as a combination of elements from different bases, e.g. a combination of B-splines with polynomials. $\qquad\square$

The next condition imposes moment conditions on the structural errors and regressors, which lead to asymptotic normality and that allow us to invoke self-normalized moderate deviation results in Jing, Shao, and Wang (2003) which were first used in the non-Gaussian analysis of Lasso in Belloni, Chen, Chernozhukov, and Hansen (2010).

Condition SM. *(i) The disturbances ζ_i and v_i have conditional variance $(\mathrm{E}[\zeta_i^2|x_i]$ and $\mathrm{E}[v_i^2|x_i])$ that are bounded uniformly from above and away from zero, uniformly in i and n. (ii) $\bar{\mathrm{E}}[|v_i|^q]$, $\bar{\mathrm{E}}[|\zeta_i|^q]$ and $\bar{\mathrm{E}}[|d_i|^q]$ are bounded uniformly in n, for some constant $q > 4$. (iii) The moments $\bar{\mathrm{E}}[|x_{ij}\zeta_i|^3]$ and $\bar{\mathrm{E}}[|x_{ij}v_i|^3]$ are bounded uniformly in $1 \leqslant j \leqslant p$, uniformly in n. The following growth conditions hold: $\log^3 p = o(n)$ and $n^{2/q}s \log(p \vee n) = o(n)$.*

3.2. **The Main Result.** The following is the main result of this paper. It shows that the post-double selection estimator is root-n consistent and is asymptotically normal. Under homoscedasticity this estimator achieves the semi-parametric efficiency bound. The plug-in estimates of the standard errors are consistent. Before stating the results, we note that the

theorem relies on the regularity conditions stated above as well as on a more technical Condition RF, which we chose to state below in the next subsection, since it is only needed for getting the standard rates for a feasible Lasso estimator under heteroskedasticity.

Theorem 1 (Estimation and Inference on Treatment Effects). *Suppose conditions ASTE(i-iv), SM and RF for (2.1) and (2.2) hold and that condition SE holds with probability approaching 1 as n grows. The post-double-Lasso estimator $\check{\alpha}$ obeys*

$$([\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) = N(0,1) + o_P(1).$$

Moreover, if Condition ASTE(v) also holds, the result continues to apply if $\bar{\mathrm{E}}[v_i^2]$ and $\bar{\mathrm{E}}[v_i^2\zeta_i^2]$ are replaced by $\mathbb{E}_n[\widehat{v}_i^2]$ and $\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2]$ for $\widehat{\zeta}_i := [y_i - d_i\check{\alpha} - x_i'\check{\beta}]\{n/(n-\widehat{s}-1)\}^{1/2}$ and $\widehat{v}_i := d_i - x_i'\widehat{\beta}$, $i = 1, \ldots, n$ where $\widehat{\beta} \in \arg\min_\beta \mathbb{E}_n[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}$.

Comment 3.3. By exploiting both equations (2.3) and (2.4) for the selection of the model, the post double selection estimator generate additional robustness as compared to selection procedures based on a single equations. The end result is that the regularity conditions appear quite weak, in particular they essentially encompass the standard regularity conditions of the kind given in Donald and Newey (2001). Robustness is also reflected in the fact that Theorem 1 permits the data-generating process (dgp) to change with n. Thus conclusions of the theorem are valid for a wide variety of sequences of dgps, and this implicitly defines the regions of uniform validity of the procedure. The regions of uniform validity appear to be substantial, which translates into good finite-sample performance of the method, as we document in the Monte-Carlo experiments reported in Section 5. □

3.3. **Auxiliary Results on Model Selection via Lasso and Post-Lasso.** The post double selection estimator applies the least squares estimator to the union of models selected via feasible Lasso. Therefore model selection properties of feasible Lasso as well properties of least square estimates for m and g based on the selected model play an important role in the derivation.

Note that either of the regression models (2.3)-(2.4) are of the following approximately sparse form:

Condition ASM. *We have data $\{(\tilde{y}_i, \tilde{z}_i, \tilde{x}_i = P(\tilde{z}_i)) : 1 \leqslant i \leqslant n\}$ consisting of i.n.i.d vectors that obey the regression model for each n:*

$$\tilde{y}_i = f(\tilde{z}_i) + \epsilon_i = \tilde{x}_i'\beta_0 + r_i + \epsilon_i,$$

$$\mathrm{E}[\epsilon_i \mid x_i] = 0, \bar{\mathrm{E}}[\epsilon_i^2] = \sigma^2,$$

$$\|\beta_0\|_0 \leqslant s, \ \mathbb{E}_n[r_i^2] \lesssim_P \sigma^2 s/n.$$

In this section we discuss the model selection properties of feasible Lasso, and derive the properties of the least squares fit to the function $f(\tilde{z}_i)$. Let $\widehat{T}$ denote the model selected by the feasible Lasso estimator $\widehat{\beta}$. Formally, set

$$\widehat{T} = \text{support}(\widehat{\beta}) = \{j \in \{1, \dots, p\} : |\widehat{\beta}_j| > 0\},$$

and define the Post-Lasso estimator $\widetilde{\beta}$ as

(3.19) $$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \ \mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2] \ : \ \beta_j = 0 \text{ for each } j \notin \widehat{T},$$

In words, the estimator is ordinary least squares applied to the data after removing the regressors that were not selected by Lasso.

We shall impose the following technical regularity conditions to deal with possibly non-Gaussian, heteroscedastic errors.

Condition RF. *(i) The following growth conditions hold* $\log^{1/3} p = o(n)$ *and* $s \log(p \vee n) = o(n)$. *(ii) The moments* $\bar{\mathrm{E}}[\tilde{y}_i^8]$ *and* $\bar{\mathrm{E}}[\epsilon_i^8]$ *are bounded uniformly in* n. *(iii) The regressors* x_i *obey:* $\max_{1 \leqslant j \leqslant p} \mathbb{E}_n[\tilde{x}_{ij}^8] \lesssim_{\mathrm{P}} 1$ *and* $\max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p} |\tilde{x}_{ij}^2| \frac{s \log(p \vee n)}{n} \to_{\mathrm{P}} 0$. *(iv) The moments* $\bar{\mathrm{E}}[\tilde{x}_{ij}^2 \epsilon_i^2]$ *are bounded away from zero and from above uniformly in* $1 \leqslant j \leqslant p$, *uniformly in* n, *and the moments* $\bar{\mathrm{E}}[\tilde{x}_{ij}^6 \tilde{y}_i^6]$ *and* $\bar{\mathrm{E}}[\tilde{x}_{ij}^6 \epsilon_i^6]$ *are bounded, uniformly in* $1 \leqslant j \leqslant p$, *uniformly in* n.

The main auxiliary result is as follows.

Lemma 1 (Model Selection Properties of Lasso and Properties of Post-Lasso). *Suppose that conditions ASM and RF hold, and that Condition SE holds for* $\mathbb{E}_n[\tilde{x}_i \tilde{x}_i']$ *with probability going to 1. Then the data-dependent model* $\widehat{T}$ *selected by Feasible Lasso estimator satisfies*

$$\widehat{s} = |\widehat{T}| \lesssim_P s,$$

and

$$\min_{\beta \in \mathbb{R}^p : \beta_j = 0 \forall j \notin \widehat{T}} \sqrt{\mathbb{E}_n[f(\tilde{z}_i) - \tilde{x}_i'\beta]^2} \lesssim_P \sigma \sqrt{\frac{s \log(p \vee n)}{n}}.$$

The Post-Lasso estimator obeys

$$\sqrt{\mathbb{E}_n[f(\tilde{z}_i) - \tilde{x}_i'\widetilde{\beta}]^2} \lesssim_P \sigma \sqrt{\frac{s \log(p \vee n)}{n}}.$$

and

(3.20) $$\|\widetilde{\beta} - \beta_0\| \lesssim_P \sqrt{\mathbb{E}_n[\{\tilde{x}_i'\widetilde{\beta} - \tilde{x}_i'\beta_0\}^2]} \lesssim_P \sigma \sqrt{\frac{s \log(p \vee n)}{n}}.$$

Comment 3.4. Thus Lasso selects a model $\widehat{T}$ that provides high-quality, near-optimal approximation to the regression function $f(\tilde{z}_i)$. The optimal approximation in our context means the approximation error of size $\sqrt{s/n}$, and here we are getting the the additional factor $\sqrt{\log(p \vee n)}$ in the rate, which is the price of not knowing the "the best" approximating model

$$T = \text{support}(\beta_0).$$

Note that Lasso generally does not recover T perfectly, that is $\widehat{T} \neq T$ in general. Moreover, no estimator can recover T perfectly in general, unless the non-zero coefficients β_0 are separated away from zero very strongly (by some sort of miracle) which seems unlikely in many econometric applications of interest. However, we do not require that; all it matters is that the selected model $\widehat{T}$ can approximate the regression function well, and the size of the model $\widehat{s} = |\widehat{T}|$ is of the same stochastic order as $s = |T|$. These are the crucial properties that we need.   □

Comment 3.5. The theorem above shows that Feasible Post-Lasso achieves the same near-oracle rate as Feasible Lasso. Notably, this occurs despite the fact that Feasible Lasso may in general fail to correctly select the oracle model T as a subset, that is $T \not\subseteq \widehat{T}$. The intuition for this result is that any components of T that Feasible Lasso misses are very unlikely to be important or their contribution can be captured by the other selected components. Lemma 1 was derived in Belloni, Chen, Chernozhukov, and Hansen (2010) for Lasso and a simple extension of Belloni, Chernozhukov, and Wang (2010) yields the result for Square-root Lasso (which could also be iterative with loadings). Similar results have been shown before for ℓ_1-penalized quantile regression (Belloni and Chernozhukov 2011a), and can be derived for other methods that yield sparse estimators. In the Gaussian context the result above was derived in Belloni, Chernozhukov, and Hansen (2010).   □

4. Generalizations and Extensions

4.1. **Split Sample Double Selection Estimator.** In this section we discuss a variant of the double selection estimator based on split sample. The underlying motivation is to attempt to reduce the possibly substantive requirement $s^2 \log^2(p \vee n) = o(n)$ that is assumed in the full-sample counterpart to the milder condition

$$s \log(p \vee n) = o(n).$$

To define the estimator divide the sample random into (approximately) equal parts a and b, with sizes $n_a = \lceil n/2 \rceil$ and $n_b = n - n_a$. (The superscripts a and b are used for variables in the

first and second subsample respectively. We typically index the subsample by $k = a, b$ and let $k^c = \{a, b\} \setminus \{k\}$.)

For each of the subsamples we apply the double selection method to select the set of controls $\widehat{I}^a := \widehat{I}_1^a \cup \widehat{I}_2^a \cup \widehat{I}_3^a$ and $\widehat{I}^b := \widehat{I}_2^b \cup \widehat{I}_2^b \cup \widehat{I}_3^b$. Then we form the double selection estimates in the two subsamples

$$(\check{\alpha}^a, \check{\beta}^a) = \operatorname*{argmin}_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \{\mathbb{E}_{n_a}[(y_{1i} - d_i \alpha - x_i'\beta)^2] \ : \ \beta_j = 0, \forall j \notin \widehat{I}^b\}, \text{and}$$

$$(\check{\alpha}^b, \check{\beta}^b) = \operatorname*{argmin}_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \{\mathbb{E}_{n_b}[(y_{1i} - d_i \alpha - x_i'\beta)^2] \ : \ \beta_j = 0, \forall j \notin \widehat{I}^a\}.$$

For an index i in the subsample k, we define the residuals $\widehat{\zeta}_i := [y_i - d_i \check{\alpha}_k - x_i'\check{\beta}_k]\{n_k/(n_k - \widehat{s}_k - 1)\}^{1/2}$ and $\widehat{v}_i := d_i - x_i'\widehat{\beta}_k$ where $\widehat{\beta}_k \in \operatorname{arg\,min}_\beta \{\mathbb{E}_{n_k}[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}^{k^c}\}$.

Importantly, the model $\widehat{I}^a$ selected based on the subsample a is the model used to fit the subsample b (and vice-versa). Finally, we combine the estimates into the split-sample double selection estimator

$$(4.21) \qquad \check{\alpha}_{ab} = \{(n_a/n)\Upsilon^a + (n_b/n)\Upsilon^b\}^{-1}((n_a/n_b)\Upsilon^a \check{\alpha}_a + (n_b/n)\Upsilon^b \check{\alpha}_b),$$

where $\Upsilon^k = D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}D^k/n_k$, $k = a, b$.

We state below sufficient regularities conditions for the analysis of the split sample double selection method.

Condition ASTESS. *(i) For each n, the data array $D_n = \{(y_{1i}, d_i, z_i), i = 1, ..., n\}$ is a sequence of i.n.i.d vectors that obey the model (2.1)-(2.2) for each n, and $x_i = P(z_i)$ is a dictionary of transformations of z_i. We allow the law of data to change with n and so all parameters can depend on n. (ii) The parameter value α_0 is bounded uniformly in n. (iii) Functions m and g admit an approximately sparse form, with sparsity index s, namely there exists $s \geqslant 1$ and β_{m0} and β_{g0} such that*

$$(4.22) \qquad m(z_i) = x_i'\beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{mi}^2]\}^2 \lesssim_P \bar{\sigma}\sqrt{s/n},$$

$$(4.23) \qquad g(z_i) = x_i'\beta_{g0} + r_{gi}, \quad \|\beta_{g0}\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{gi}^2]\}^2 \lesssim_P \bar{\sigma}\sqrt{s/n}.$$

(iv) The sparsity index obeys $s\log(p \vee n) = o(n)$. (v) For subsamples $k = a, b$, the size of the amelioration set obeys $\widehat{s}_3^k \lesssim_P 1 \vee \widehat{s}_1^k \vee \widehat{s}_2^k$, $|m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k| = o_P(\sqrt{n_k})$ where $\mathcal{M}_{\widehat{I}^{k^c}}$ is the orthogonal projection operator associated with the covariates in $\widehat{I}^{k^c}$, $k^c = \{a, b\} \setminus \{k\}$. (vi) $\mathbb{E}_{nk}[(v_i + r_{mi})^2(\zeta_i + r_{gi})^2] - \bar{\mathbb{E}}_k[v_i^2\zeta_i^2] \to_P 0$, $\max_{i \leqslant n} \|(r_{gi}, r_{mi}, \zeta_i, v_i, \widehat{\zeta}_i, \widehat{v}_i)'\|_\infty^2 s\log(p \vee n) = o_P(n)$.

The Conditions ASTESS(i)-(iv) agree with the corresponding conditions in ASTE. The remaining conditions ASTESS(v)-(vi) are implied by Condition ASTE. We note that Condition ASTESS(vi) is needed only for obtaining consistent estimates of the asymptotic variance. Such conditions are mild since they do not require uniform estimation of the functions g and m.

The next result establishes that the split-sample double selection estimator $\widehat{\alpha}_{ab}$ has the similar large sample properties as the (full-sample) double selection estimator under under weaker growth condition.

Theorem 2 (Inference on Treatment Effects, Split Sample). *Suppose conditions ASTESS(i-v), SM and RF for (2.1) and (2.2) hold and that condition SE holds with probability approaching 1 as n grows for each subsample. The split sample post-double-selection estimator $\check{\alpha}_{ab}$ obeys,*

$$([\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) = N(0,1) + o_P(1).$$

Moreover, if Condition ASTESS(vi) also holds, the result continues to apply if $\bar{\mathrm{E}}[v_i^2]$ and $\bar{\mathrm{E}}[v_i^2\zeta_i^2]$ are replaced by $\mathbb{E}_n[\widehat{v}_i^2]$ and $\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2]$.

5. Monte-Carlo Example

In this section, we compare the estimation strategies proposed above in the following model:

$$(5.24) \qquad y_i = d_i'\alpha_0 + x_i'\beta_0 + \zeta_i, \quad \zeta_i \sim N(0,\sigma_\zeta^2)$$

where the covariates $x \sim N(0,\Sigma)$, $\Sigma_{kj} = (0.5)^{|j-k|}$, and

$$(5.25) \qquad d_i = \tilde{x}_i'\eta_0 + v_i, \quad v_i \sim N(0,\sigma_v^2)$$

with $\sigma_\zeta = \sigma_v = 1$, and $\sigma_{\zeta v} = 0$. The dimension p of the covariates x is 200, and the sample size n is 100. We set $\alpha_0 = 1$ and

$$\beta_0 = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5},\ 0,\ 0,\ 0,\ 0,\ 0,\ 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, \ldots, 0\right)',$$

$$\eta_0 = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}, \frac{1}{10},\ 0, \ldots\ldots\ldots\ldots\ldots, 0\right)'.$$

We set λ according with $1 - \gamma = .95$. For each repetition we draw new x's, ζ's and v's.

We compare the proposed post double selection method with the following approaches: Lasso, estimate α_0 by applying a Feasible Lasso method to model (2.1) without penalizing α; Post-single selection 1, estimate α_0 by applying a Feasible Post-Lasso method to model (2.1) without penalizing α; Post-single selection 2, estimate α_0 by applying least squares regression of y on d and control terms selected by a Feasible Lasso regression of d_i on x_i in (2.2); Oracle,

estimate α_0 by applying least squares regression of y on d and control terms in the true support of β_0 (which is unavailable outside the experiment).

We summarize the inference performance of these methods in Table 1 which illustrates mean bias, standard deviation, and rejection probabilities of 95% confidence intervals. As we had expected, Lasso and Post-single selection 1 exhibit a large mean bias which dominates the estimation error and results in poor performance of conventional inference methods. On the other hand, the Post-single selection 2 has a small bias relative to estimation error but is substantially more variable than Post-double selection and produces a conservative test, a test with size much smaller than the nominal level. Notably, the Post-double selection provides coverage that is close to the promised 5% level and has the smallest mean bias and standard deviation.

Partial Linear Model Simulation Results

Estimator	Mean Bias	Std. Dev.	rp(0.05)
Baseline			
Lasso	0.644	0.093	1.000
Post-single selection 1	0.415	0.209	0.877
Post-single selection 2	0.0908	0.194	0.004
Oracle	-0.0003	0.100	0.044
Our proposal			
Post-double selection	-0.0041	0.111	0.054

TABLE 1. Results are based on 1000 simulation replications of the partially linear model (5.24) where $p = 200$ and $n = 100$. We report mean bias (Mean Bias), standard deviation (Std. Dev.), and rejection frequency for 5% level tests (rp(.05)) for the four estimators described in Section 7.1.

6. Empirical Example: Estimating the Effect of Abortion on Crime

In the preceding sections, we have provided results demonstrating how variable selection methods, focusing on the case of Lasso-based methods, can be used to estimated treatment effects in models in which we believe the variable of interest is exogenous conditional on observables. We further illustrate the use of these methods in this section by reexamining Levitt and Donohue's (2001) study of the impact of abortion on crime rates. In the following, we briefly review Donohue III and Levitt (2001) and then present estimates obtained using the methods developed in this paper.

Donohue III and Levitt (2001) discuss two key arguments for a causal channel relating abortion to crime. The first is simply that more abortion among a cohort results in an otherwise smaller cohort and so crime 15 to 25 years later, when this cohort is in the period when its members are most at risk for committing crimes, will be otherwise lower given the smaller cohort size. The second argument is that abortion gives women more control over the timing of their fertility allowing them to more easily assure that childbirth occurs at a time when a more favorable environment is available during a child's life. For example, access to abortion may make it easier to ensure that a child is born at a time when the family environment is stable, the mother is more well-educated, or household income is stable. This second channel would mean that more access to abortion could lead to lower crime rates even if fertility rates remained constant.

The basic problem in estimating the causal impact of abortion on crime is that state-level abortion rates are not randomly assigned, and it seems likely that there will be factors that are associated to both abortion rates and crime rates. It is cleat that any association between the current abortion rate and the current crime rate is likely to be spurious. However, even if one looks at say the relationship between the abortion rate 18 years in the past and the crime rate among current 18 year olds, the lack of random assignment makes establishing a causal link difficult without adequate controls. An obvious confounding factor is the existence of persistent state-to-state differences in policies, attitudes, and demographics that are likely related to the overall state level abortion and crime rates. It is also important to control for flexibly for aggregate trends. For example, it could be the case that national crime rates were falling over this period while national abortion rates were rising but that these trends were driven by completely different factors. Without controlling for these trends, one would mistakenly associate the reduction in crime to the increase in abortion. In addition to these overall differences across states and times, there are other time varying characteristics such as state-level income, policing, or drug-use to name a few that could be associated with current crime and past abortion.

To address these confounds, Donohue III and Levitt (2001) estimate a model with annual state-level data with crime rate data running from 1985 to 1997 in which they condition on a number of these factors. Their basic specification is

$$(6.26) \qquad\qquad y_{cit} = \alpha a_{cit} + w'_{it}\beta + \delta_i + \gamma_t + \varepsilon_{it}$$

where i indexes states, t indexes times, $c \in \{\text{violent, property, murder}\}$ indexes type of crime, δ_i are state-specific effects that control for any time-invariant state-specific characteristics, γ_t are time-specific effects that control flexibly for any aggregate trends, w_{it} are a set of

control variables to control for time-varying confounding state-level factors, a_{cit} is a measure of the abortion rate relevant for type of crime c,[6] and y_{cit} is the crime-rate for crime type c. Donohue III and Levitt (2001) use the log of lagged prisoners per capita, the log of lagged police per capita, the poverty rate, AFDC generosity at time $t-15$, a dummy for concealed weapons law, and beer consumption per capita for w_{it}, the set of time-varying state-specific controls. Tables IV and V in Donohue III and Levitt (2001) present baseline estimation results based on (6.26) as well as results from different models which vary the sample and set of controls to show that the baseline estimates are robust to small deviations from (6.26). We refer the reader to the original paper for additional details, data definitions, and institutional background.

For our analysis, we take the argument that the abortion rates defined above may be taken as exogenous relative to crime rates once observables have been conditioned on from Donohue III and Levitt (2001) as given. Given the seemingly obvious importance of controlling for state and time effects, we account for these in all models we estimate. We choose to eliminate the state effects via differencing rather than including a full set of state dummies but include a full set of time dummies in every model. Thus, we will estimate models of the form

$$(6.27) \qquad y_{cit} - y_{cit-1} = \alpha(a_{cit} - a_{cit-1}) + z'_{it}\kappa + \gamma_t + \eta_{it}.$$

We use the same state-level data as Donohue III and Levitt (2001) but delete Alaska, Hawaii, and Washington, D.C. which gives a sample with 48 cross-sectional observations and 12 time series observations for a total of 576 observations. With these deletions, our baseline estimates using the same controls as in (6.26) are quite similar to those reported in Donohue III and Levitt (2001). Baseline estimates from Table IV of Donohue III and Levitt (2001) and our baseline estimates based on the differenced version of (6.26) are given in the first and second row of Table 2 respectively.

Our main point of departure from Donohue III and Levitt (2001) is that we allow for a much richer set z_{it} than allowed for in x_{it} in model (6.26). Our z_{it} includes higher-order terms and interactions of the control variables defined above. In addition, we put initial conditions and initial differences of x_{it} and a_{it} into our vector of controls z_{it}. This addition allows for the

---

[6]This variable is constructed as weighted average of abortion rates where weights are determined by the fraction of the type of crime committed by various age groups. For example, if 60% of violent crime were committed by 18 year olds and 40% were committed by 19 year olds in state i, the abortion rate for violent crime at time t in state i would be constructed as .6 times the abortion rate in state i at time $t-18$ plus .4 times the abortion rate in state i at time $t-19$. See Donohue III and Levitt (2001) for further detail and exact construction methods.

possibility that there may be some feature of a state that is associated both with its growth rate in abortion and its growth rate in crime. For example, having an initially high-levels of abortion could be associated with having high-growth rates in abortion and low growth rates in crime. Failure to control for this factor could then lead to misattributing the effect of this initial factor, perhaps driven by policy or state-level demographics, to the effect of abortion. Finally, we allow for more general trends by allowing for an aggregate quadratic trend in z_{it} as well as interactions of this quadratic trend with control variables. This gives us a set of 251 control variables to select among in addition to the 12 time effects that we include in every model.[7]

Note that interpreting estimates of the effect of abortion from model (6.26) as causal relies on the belief that there are no higher-order terms of the control variables, no interaction terms, and no additional excluded variables that are associated both to crime rates and the associated abortion rate. Thus, controlling for a large set of variables as described above is desirable from the standpoint of making this belief more plausible. At the same time, naively controlling lessens our ability to identify the effect of interest and thus tends to make estimates far less precise. The effect of estimating the abortion effect conditioning on the full set of 251 potential controls described above is given in the third row of Table 2. As expected, all coefficients are estimated very imprecisely. Of course, very few researchers would consider using 251 controls with only 576 observations due to exactly this issue.

We are faced with a tradeoff between controlling for very few variables which may leave us wondering whether we have included sufficient controls for the exogeneity of the treatment and controlling for so many variables that we are essentially mechanically unable to learn about the effect of the treatment. The variable selection methods developed in this paper offer one resolution to this tension. The assumed sparse structure maintains that there is a small enough set of variables that one could potentially learn about the treatment but adds substantial flexibility to the usual case where a researcher considers only a few control variables by allowing this set to be found by the data from among a large set of controls. Thus, the approach should complement the usual careful specification analysis by providing a researcher an efficient, data-driven way to search for a small set of influential confounds from among a sensibly chosen broad set of potential confounding variables.

---

[7]The exact identities of the 251 potential controls is available upon request. It consists of linear and quadratic terms of each continuous variable in w_{it}, interactions of every variable in w_{it}, initial levels and initial differences of w_{it} and a_{it}, and interactions of these variables with a quadratic trend.

In the abortion example, we use the post-double-selection estimator defined in Section 2.2 for each of our dependent variables. For violent crime, ten variables are selected in the abortion equation,[8] and one is selected in the crime equation.[9] For property crime, eight variables are selected in the abortion equation,[10] and six are selected in the crime equation.[11] For murder, eight variables are selected in the abortion equation,[12] and none were selected in the crime equation.

Estimates of the causal effect of abortion on crime obtained by searching for confounding factors among our set of 251 potential controls are given in the fourth row of Table 2. Each of these estimates is obtained from the least squares regression of the crime rate on the abortion rate and the 11, 14, and eight controls selected by the double-Lasso procedure for violent crime, property crime, and murder respectively. The estimates for the effect of abortion on violent crime and the effect of abortion on murder are quite imprecise, producing 95% confidence intervals that encompass large positive and negative values. The estimated effect for property crime is roughly in line with the previous estimates though it is no longer significant at the 5% level but is significant at the 10% level. Note that the double-Lasso produces models that are not of vastly different size than the "intuitive" model (6.26), though it does produce a larger model in each case.

It is very interesting that one would draw qualitatively different conclusions from the estimates obtained using formal variable selection than from the estimates obtained using a small set of intuitively selected controls. Looking at the set of selected control variables, we see that initial conditions and interactions with trends are selected across all dependent variables. The selection of this set of variables suggests that there are initial factors which are associated with

---

[8]The selected variables are AFDC generosity squared, beer consumption squared, the initial poverty change, initial income, initial income squared, the initial change in prisoners per capita squared interacted with the trend, initial income interacted with the trend, the initial change in the abortion rate, the initial change in the abortion rate interacted with the trend, and the initial level of the abortion rate.

[9]The initial level of the abortion rate interacted with time is selected.

[10]The selected variables are income, the initial poverty change, the initial change in prisoners per capita squared, the initial level of prisoners per capita, initial income, the initial change in the abortion rate, the initial change in the abortion rate interacted with the trend, and the initial level of the abortion rate.

[11]The six variables are the initial level of AFDF generosity, the initial level of income interacted with the trend and the trend squared, the initial level of income squared interacted with the trend and the trend squared, and the initial level of the abortion rate interacted with the trend.

[12]The selected variables are AFDC generosity, beer consumption squared, the change in beer consumption squared, the change in beer consumption squared times the trend and the trend squared, initial income times the trend, the initial change in the abortion rate interacted with the trend, and the initial level of the abortion rate.

the change in the abortion rate. We also see that we cannot precisely determine the effect of the abortion rate on crime rates once one accounts for initial conditions. Of course, this does not mean that the effects of the abortion rate provided in the first two rows of Table 2 are not representative of the true causal effects. It does, however, imply that this conclusion is strongly predicated on the belief that there are not other unobserved state-level factors that are correlated to both initial values of the controls and abortion rates, abortion rate changes, and crime rate changes.

	Violent Crime		Property Crime		Murder	
	$\widehat{\alpha}$	Std. Err.	$\widehat{\alpha}$	Std. Err.	$\widehat{\alpha}$	Std. Err.
Donohue III and Levitt (2001) Table IV	-0.129	0.024	-0.091	0.018	-0.121	0.047
First-Difference	-0.152	0.034	-0.108	0.022	-0.204	0.068
All Controls	0.294	0.472	-0.068	0.157	0.321	1.109
Post-Double-Selection	-0.087	0.181	-0.094	0.051	0.006	0.280

TABLE 2. The table displays the estimated coefficient on the abortion rate, $\widehat{\alpha}$ and its estimated standard error. Numbers in the first row are taken from Donohue and Levitt (2001) Table IV, columns (2), (4), and (6). The remaining rows are estimated by first differences, include a full set of time dummies, and use standard errors clustered at the state-level. Estimates in the row labeled "First-Difference" are obtained using the same controls as in the first row. Estimates in the row labeled "All Controls" use 251 control variables as discussed in the text. Estimates in the row "Post-Double-Selection" use the variable selection technique developed in this paper to search among the set of 251 potential controls.

We believe that the example in this section illustrates how one may use modern variable selection techniques to complement causal analysis in economics. In the abortion example, we are able to search among a large set of controls and transformations of variables when trying to estimate the effect of abortion on crime. Considering a large set of controls makes the underlying assumption of exogeneity of the abortion rate conditional on observables more plausible, while the methods we develop allow us to produce an end-model which is of manageable dimension. Interestingly, we see that one would draw quite different conclusions from the estimates obtained using formal variable selection. Looking at the variables selected, we can also see that this change in interpretation is being driven by the variable selection method's selecting different variables, specifically initial values of the abortion rate and controls, than are usually considered. Thus, it appears that the usual interpretation hinges on the prior belief that initial values should be excluded from the structural equation.

Appendix A. Iterated Estimation of Penalty Loadings

In the case of Lasso under heteroskedasticity, the penalty loadings (2.10) require the practitioner to fill in a their values. Theoretically, any upper bound on l_j's but in various examples we found that this approach leads to overpenalization. Here we briefly discuss iterative procedures to estimate l_j's similar to the ones described in Belloni and Chernozhukov (2011b). Let I_0 be a set of regressors that is included in the model. Note that I_0 is always non-empty since it will always include the intercept. Let $\bar{\beta}(I_0)$ be the least squares estimator of the coefficients on the covariates associated with I_0, and define $\widehat{l}_{jI_0} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\bar{\beta}(I_0))^2]}$.

An algorithm for estimating the loadings using Lasso is as follows:

Algorithm 1 (Estimation of Lasso loadings using Lasso iterations). *Set $\widehat{l}_{j,0} := \widehat{l}_{jI_0}$, $j = 1, \ldots, p$. Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Lasso estimator $\widehat{\beta}$ based on the loadings $\widehat{l}_{j,k}$. (2) Set $\widehat{l}_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}$. (3) If $\max_{1 \leqslant j \leqslant p} |\widehat{l}_{j,k} - \widehat{l}_{j,k+1}| \leqslant \nu$ or $k > K$, report $\widehat{l}_{j,k+1}$, $j = 1, \ldots, p$; otherwise set $k \leftarrow k + 1$ and go to (1).*

Similarly, an algorithm for estimating σ using Post-Lasso is as follows:

Algorithm 2 (Estimation of Lasso loadings using Post-Lasso iterations). *Set $\widehat{l}_{j,0} := \widehat{l}_{jI_0}$, $j = 1, \ldots, p$. Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Post-Lasso estimator $\widetilde{\beta}$ based on the loadings $\widehat{l}_{j,k}$. (2) For $\widehat{s} = \|\widetilde{\beta}\|_0 = |\widehat{T}|$ set $l_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}\sqrt{n/(n - \widehat{s})}$. (3) If $\max_{1 \leqslant j \leqslant p} |\widehat{l}_{j,k} - \widehat{l}_{j,k+1}| \leqslant \nu$ or $k > K$, report $\widehat{l}_{j,k+1}$, $j = 1, \ldots, p$; otherwise, set $k \leftarrow k + 1$ and go to (1).*

To estimate the loadings in the case of Square-root Lasso one can proceed similarly by setting $\widehat{l}_{j,0} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\bar{\beta}(I_0))^2]}/\sqrt{\mathbb{E}_n[(y_i - x_i'\bar{\beta}(I_0))^2]}$. Nonetheless, the self normalization allows for the alternative initial proposal $\widehat{l}_{j,0} := 2\{\mathbb{E}_n[x_{ij}^4]\}^{1/4}$, $j = 1, \ldots, p$. Such choice is valid, $l_{j,0} + o_P(1) \leqslant \widehat{l}_{j,0}$ uniformly in $j = 1, \ldots, p$), for a broad class of marginal distributions for ϵ_i that include all t-distributions with degree of freedom greater than five. The algorithm below can be applied with either of these choices.

Algorithm 3 (Estimation of Square-root Lasso loadings using Square-root Lasso iterations). *Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Square-root Lasso estimator $\widehat{\beta}$ based on the loadings $\widehat{l}_{j,k}$. (2) Set $\widehat{l}_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat{\beta})^2]}/\sqrt{\mathbb{E}_n[(y_i - x_i'\widehat{\beta})^2]}$. (3) If*

$\max_{1 \leqslant j \leqslant p} |\widehat{l}_{j,k} - \widehat{l}_{j,k+1}| \leqslant \nu$ *or $k > K$, report $\widehat{l}_{j,k+1}$, $j = 1, \ldots, p$; otherwise set $k \leftarrow k + 1$ and go to (1).*

Algorithm 4 (Estimation of Square-root Lasso loadings using Post-Square-root Lasso iterations)**.** *Set $k = 0$, and specify a small constant $\nu \geqslant 0$ as a tolerance level and a constant $K > 1$ as an upper bound on the number of iterations. (1) Compute the Post-Square-root Lasso estimator $\widetilde{\beta}$ based on the loadings $\widehat{l}_{j,k}$. (2) Set $\widehat{l}_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widetilde{\beta})^2]}/\sqrt{\mathbb{E}_n[(y_i - x_i'\widetilde{\beta})^2]}$. (3) If $\max_{1 \leqslant j \leqslant p} |\widehat{l}_{j,k} - \widehat{l}_{j,k+1}| \leqslant \nu$ or $k > K$, report $\widehat{l}_{j,k+1}$, $j = 1, \ldots, p$; otherwise set $k \leftarrow k + 1$ and go to (1).*

APPENDIX B. PROOF OF THEOREM 1

We use the standard matrix notation, namely $Y_1 = [y_{11}, ..., y_{1n}]'$, $X = [x_1, ..., x_n]'$, $D = [d_1, ..., d_n]'$, $V = [v_1, ..., v_n]'$, $\zeta = [\zeta_1, ..., \zeta_n]'$, $m = [m_1, ..., m_n]'$, $R_m = [r_{m1}, ..., r_{mn}]'$, $g = [g_1, ..., g_n]'$, $R_g = [r_{g1}, ..., r_{gn}]'$, and so on. For $A \subset \{1, ..., p\}$, let $X[A] = \{X_j, j \in A\}$, where $\{X_j, j = 1, ..., p\}$ are the columns of X. Let

$$\mathcal{P}_A = X[A](X[A]'X[A])^- X[A]'$$

be the projection operator sending vectors in $\mathbb{R}^n$ onto $\mathrm{span}[X[A]]$, and let $\mathcal{M}_A = \mathrm{I}_n - \mathcal{P}_A$ be the projection onto the subspace that is orthogonal to $\mathrm{span}[X[A]]$. For a vector $Z \in \mathbb{R}^n$, let

$$\tilde{\beta}_Z(A) := \arg\min_{b \in \mathbb{R}^p} \|Z - X'b\|^2 : \ b_j = 0, \ \forall j \notin A,$$

be the coefficient of linear projection of Z onto $\mathrm{span}[X[A]]$. If $A = \varnothing$, interpret $\mathcal{P}_A = 0_n$, and $\tilde{\beta}_Z = 0_p$.

Finally, denote $\phi_{\min}(m) = \phi_{\min}(m)[\mathbb{E}_n[x_i x_i']]$ and $\phi_{\max}(m) = \phi_{\max}(m)[\mathbb{E}_n[x_i x_i']]$.

Step 1.(Main) Write $\check{\alpha} = [D'\mathcal{M}_{\widehat{I}}D/n]^{-1} [D'\mathcal{M}_{\widehat{I}}Y_1/n]$ so that

$$\sqrt{n}(\check{\alpha} - \alpha_0) = [D'\mathcal{M}_{\widehat{I}}D/n]^{-1} [D'\mathcal{M}_{\widehat{I}}(g + \zeta)/\sqrt{n}] =: ii^{-1} \cdot i.$$

By Steps 2 and 3, $ii = V'V/n + o_P(1)$ and $i = V'\zeta/\sqrt{n} + o_P(1)$. Next note that $V'V/n = \mathrm{E}[V'V/n] + o_P(1)$ by Chebyshev, and because $\mathrm{E}[V'V/n]$ are bounded from above and away from zero by assumption, we have $ii^{-1} = \mathrm{E}[V'V/n]^{-1} + o_P(1)$.

Letting $\Gamma = \mathrm{diag}(\zeta_1^2, \ldots, \zeta_n^2)$, define

$$Z_n = (\mathrm{E}[V'V/n]^{-1}\mathrm{E}[V'\Gamma V/n]\mathrm{E}[V'V/n]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) = \mathbb{G}_n[z_{i,n}] + o_P(1),$$

where $z_{i,n} = (\mathrm{E}[V'V/n]^{-1}\mathrm{E}[V'\Upsilon V/n]\mathrm{E}[V'V/n]^{-1})^{-1/2}v_i\zeta_i/\sqrt{n}$ are i.n.i.d. with mean zero. We have that for some small enough $\delta > 0$

$$\bar{\mathrm{E}}|z_{i,n}|^{2+\delta} \lesssim \bar{\mathrm{E}}\left[|v_i|^{2+\delta}|\zeta_i|^{2+\delta}\right] \lesssim \sqrt{\bar{\mathrm{E}}|v_i|^{4+2\delta}}\sqrt{\bar{\mathrm{E}}|\zeta_i|^{4+2\delta}} \lesssim 1,$$

by Condition SM.

This condition verifies the Lyapunov condition and thus implies that $Z_n \to_d N(0,1)$.

Step 2. (Behavior of i.) Decompose

$$i = V'\zeta/\sqrt{n} + \underbrace{m'\mathcal{M}_{\widehat{I}}g/\sqrt{n}}_{=:i_a} + \underbrace{m'\mathcal{M}_{\widehat{I}}\zeta/\sqrt{n}}_{=:i_b} + \underbrace{V'\mathcal{M}_{\widehat{I}}g/\sqrt{n}}_{=:i_c} - \underbrace{V'\mathcal{P}_{\widehat{I}}\zeta/\sqrt{n}}_{=:i_d}.$$

First, by Step 4 and 5 below we have

$$|i_a| = |m'\mathcal{M}_{\widehat{I}}g/\sqrt{n}| = \sqrt{n}\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\|\|\mathcal{M}_{\widehat{I}}m/\sqrt{n}\| \lesssim_P \sqrt{[s\log(p\vee n)]^2/n} = o(1)$$

where the last bound follows from the growth condition $s^2\log^2(p\vee n) = o(n)$.

Second, using decomposition $m = X\beta_{m0} + R_m$, we have

$$|i_b| \leqslant |R'_m\zeta/\sqrt{n}| + |(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})'X'\zeta/\sqrt{n}| \lesssim_P \sqrt{[s\log(p\vee n)]^2/n} = o_P(1),$$

where $|R'_m\zeta/\sqrt{n}| \lesssim_P \sqrt{R'_m R_m/n} \lesssim_P \sqrt{s/n}$ by Chebyshev inequality and by assumption ASTE(iii), and

$$|(\tilde{\beta}_m(\widehat{I}) - \beta_{m0})'X'\zeta/\sqrt{n}| \leqslant \|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\|_1\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{[s^2\log(p\vee n)]/n}\sqrt{\log(p\vee n)},$$

$\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\|_1 \leqslant \sqrt{\widehat{s}}\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\| \lesssim_P \sqrt{[s^2\log(p\vee n)]/n}$ by Step 4, using that $\widehat{s} \lesssim_P s$ by Lemma 1, $\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}$ by Lemma 7 in Belloni, Chen, Chernozhukov, and Hansen (2010) under SM. Third, using similar reasoning, decomposition $g = X\beta_{g0} + R_g$, and Step 5, conclude

$$|i_c| \leqslant |R'_g\zeta| + |(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})'X'V/\sqrt{n}| \lesssim_P \sqrt{[s\log(p\vee n)]^2/n} = o_P(1).$$

Fourth, using that $\widehat{s} \lesssim_P s$ by Lemma 1 so that $\phi_{\min}^{-1}(\widehat{s}) \lesssim_P 1$ by condition SE, conclude,

$$|i_d| \leqslant |\tilde{\beta}_V(\widehat{I})'X'\zeta/\sqrt{n}| \leqslant \|\tilde{\beta}_V(\widehat{I})\|_1\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{n}\sqrt{(s\log p)^2/n^2} = o_P(1),$$

since $\|\tilde{\beta}_V(\widehat{I})\|_1 \leqslant \sqrt{\widehat{s}}\|\tilde{\beta}_V(\widehat{I})\| \leqslant \sqrt{\widehat{s}}\|(X[\widehat{I}]'X[\widehat{I}])^{-1}X[\widehat{I}]'V/n\| \leqslant \sqrt{\widehat{s}}\phi_{\min}^{-1/2}(\widehat{s})\sqrt{\widehat{s}}\|X'V/\sqrt{n}\|_\infty/\sqrt{n}$
$\lesssim_P s\sqrt{[\log(p\vee n)]/n}$.

Step 3. (Behavior of ii.) Decompose

$$ii = (m+V)'\mathcal{M}_{\widehat{I}}(m+V)/n = V'V/n + \underbrace{m'\mathcal{M}_{\widehat{I}}m/n}_{=:ii_a} + \underbrace{2m'\mathcal{M}_{\widehat{I}}V/n}_{=:ii_b} - \underbrace{V'\mathcal{P}_{\widehat{I}}V/n}_{=:ii_c}.$$

Then $|ii_a| \lesssim_P [s \log(p \vee n)]/n = o_P(1)$ by Step 4, $|ii_b| \lesssim_P [s \log(p \vee n)]/n = o_P(1)$ by reasoning similar to deriving the bound for $|i_b|$, and $|ii_c| \lesssim_P [s \log(p \vee n)]/n = o_P(1)$ by reasoning similar to deriving the bound for $|i_d|$.

Step 4. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}} m\|$ and related quantities.) Note that

$$
\begin{aligned}
\|\mathcal{M}_{\widehat{I}} m\| &\leqslant \|\mathcal{M}_{\widehat{I}_1} m\| \\
&\leqslant \|X \tilde{\beta}_D(\widehat{I}_1) - m\| \\
&\leqslant \|X(\tilde{\beta}_D(\widehat{I}_1) - \beta_{m0})\| + \|R_m\|
\end{aligned}
$$

since $\widehat{I}_1 \subseteq \widehat{I}$. By Condition ASTE(iii), we have $\|R_m/\sqrt{n}\| \lesssim_P \sqrt{s/n}$ and, also using Lemma 1 so that $\widehat{s} \lesssim_P s$, we have

$$
\begin{aligned}
\|\tilde{\beta}_D(\widehat{I}_1) - \beta_{m0}\| &\leqslant \sqrt{1/\phi_{\min}(\widehat{s})} \|X(\tilde{\beta}_D(\widehat{I}_1) - \beta_{m0})/\sqrt{n}\| \\
&\lesssim_P \sqrt{[s \log(p \vee n)]/n}
\end{aligned}
$$

since $1/\phi_{\min}(\widehat{s}) \lesssim_P 1$ by condition SE. Thus we also have established

$$
\|\tilde{\beta}_m(\widehat{I}) - \beta_{m0}\| \lesssim_P \sqrt{[s \log(p \vee n)]/n}.
$$

Step 5. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}} g\|$ and related quantities.) Note that

$$
\begin{aligned}
\|\mathcal{M}_{\widehat{I}} g\| &\leqslant \|\mathcal{M}_{\widehat{I}_2} g\| \\
&\leqslant \|X \tilde{\beta}_{Y_1}(\widehat{I}_2) - g\| \\
&\leqslant \|X(\tilde{\beta}_{Y_1}(\widehat{I}_2) - \beta_{g0})\| + \|R_g\| \\
&\lesssim_P \sqrt{s \log(p \vee n)}
\end{aligned}
$$

since $\widehat{I}_2 \subseteq \widehat{I}$, the triangle inequality, $\|R_g/\sqrt{n}\| \lesssim_P \sqrt{s/n}$, and by Lemma 1, similarly to Step 4 using SE, it follows that $\|X(\tilde{\beta}_{Y_1}(\widehat{I}_2) - \beta_{g0})/\sqrt{n}\| \lesssim_P \sqrt{[s \log(p \vee n)]/n}$ and $\|\tilde{\beta}_g(\widehat{I}) - \beta_{g0}\| \lesssim_P \sqrt{[s \log(p \vee n)]/n}$.

Step 6. (Variance Estimation.) Since $\widehat{s} \lesssim_P s = o(n)$, $(n - \widehat{s} - 1)/n = o_P(1)$, so we can use n as the denominator. Hence consider

$$
\mathbb{E}_n[\widehat{v}_i^2] = D' \mathcal{M}_{\widehat{I}} D/n = V'V/n + o_P(1) = \bar{\mathrm{E}}[v_i^2] + o_P(1)
$$

by Step 3 and $\bar{\mathrm{E}}[|v_i|^q] \lesssim 1$ for some $q > 4$ by condition SM.

Let $\tilde{v}_i = v_i + r_{mi}$ and $\tilde{\zeta}_i = \zeta_i + r_{gi}$. Recall that by Condition ASTE(v) we have $\mathbb{E}_n[\tilde{v}_i^2 \tilde{\zeta}_i^2] - \bar{\mathrm{E}}[v_i^2 \zeta_i^2] \to_P 0$. To show that $\mathbb{E}_n[\widehat{v}_i^2 \widehat{\zeta}_i^2] - \mathbb{E}_n[\tilde{v}_i^2 \tilde{\zeta}_i^2] \to_P 0$ we start applying a triangular inequality

$$
|\mathbb{E}_n[\widehat{v}_i^2 \widehat{\zeta}_i^2 - \tilde{v}_i^2 \tilde{\zeta}_i^2]| \leqslant |\mathbb{E}_n[(\widehat{v}_i^2 - \tilde{v}_i^2) \tilde{\zeta}_i^2]| + |\mathbb{E}_n[\widehat{v}_i^2(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]|.
$$

Then,

$$
\begin{aligned}
|\mathbb{E}_n[\widehat{v}_i^2(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]| \quad &\leqslant 2\mathbb{E}_n[\{d_i(\alpha_0 - \check{\alpha})\}^2\widehat{v}_i^2] + 2\mathbb{E}_n[\{x_i'(\check{\beta} - \beta_{g0})\}^2\widehat{v}_i^2] \\
&\quad + |2\mathbb{E}_n[(\zeta_i + r_{gi})d_i(\alpha_0 - \check{\alpha})\widehat{v}_i^2]| + |2\mathbb{E}_n[(\zeta_i + r_{gi})x_i'(\check{\beta} - \beta_{g0})\widehat{v}_i^2]| \\
&\lesssim_P o_P(1)
\end{aligned}
$$

by the relations below.

As a consequence of Condition SM we have $\mathrm{E}[\max_{i\leqslant n} d_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i\leqslant n} \zeta_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i\leqslant n} v_i^2] \lesssim n^{2/q}$, thus by Markov inequality we have $\|D\|_\infty + \|\zeta\|_\infty + \|V\|_\infty \lesssim_P n^{1/q}$. Letting $\widehat{T}_g = \mathrm{support}(\beta_{g0}) \cup \widehat{I}$, we also have $\max_{i\leqslant n}\|x_{i\widehat{T}_g}\|^2 \leqslant |\widehat{T}_g| \max_{i\leqslant n}\|x_i\|_\infty^2 \lesssim_P s\max_{i\leqslant n}\|x_i\|_\infty^2$ by the sparsity assumption in ASTE and the sparsity bound in Lemma 1. Also by Condition ASTE(v) $(\|R_g\|_\infty + \|\zeta\|_\infty)\max_{i\leqslant n}\|x_i\|_\infty\sqrt{[s^2\log(p\vee n)]/n} = o_P(1)$ and $(\|R_g\|_\infty + \|\zeta\|_\infty)\max_{i\leqslant n}|d_i|/\sqrt{n} = o_P(1)$. Therefore, we have the following relations:

$$
\begin{aligned}
\mathbb{E}_n[\{d_i(\alpha_0 - \check{\alpha})\}^2\widehat{v}_i^2] \quad &\leqslant \|D\|_\infty^2|\alpha_0 - \check{\alpha}|^2\mathbb{E}_n[\widehat{v}_i^2] \lesssim_P n^{-1+[2/q]} = o(1) \\
\mathbb{E}_n[\{x_i'(\check{\beta} - \beta_{g0})\}^2\widehat{v}_i^2] \quad &\leqslant (\max_{i\leqslant n}\{x_i'(\check{\beta} - \beta_{g0})\}^2)\mathbb{E}_n[\widehat{v}_i^2] \lesssim_P \max_{i\leqslant n}\|x_{i\widehat{T}_g}\|^2\frac{s\log(p\vee n)}{n} \lesssim_P o(1) \\
|\mathbb{E}_n[(\zeta_i + r_{gi})d_i(\alpha_0 - \check{\alpha})\widehat{v}_i^2]| \quad &\leqslant (\|\zeta\|_\infty + \|R_g\|_\infty)\|D\|_\infty\mathbb{E}_n[\widehat{v}_i^2]|\alpha_0 - \check{\alpha}_k| = o_P(1) \\
|\mathbb{E}_n[(\zeta_i + r_{gi})x_i'(\check{\beta} - \beta_{g0})\widehat{v}_i^2]| \quad &\lesssim_P (\|\zeta\|_\infty + \|R_g\|_\infty)\max_{i\leqslant n}\|x_i\|_\infty\sqrt{s}\|\check{\beta} - \beta_{g0}\|\mathbb{E}_n[\widehat{v}_i^2] = o_P(1)
\end{aligned}
$$

since $\mathbb{E}_n[\widehat{v}_i^2] \lesssim_P 1$, $\|\check{\beta} - \beta_{g0}\|^2 \lesssim_P [s\log(p\vee n)]/n$ by Lemma 1, and $|\check{\alpha} - \alpha_0|^2 \lesssim_P 1/n$ by Step 1.

Similarly, $\mathbb{E}_n[(\widehat{v}_i^2 - v_i^2)\tilde{\zeta}_i^2] = o_P(1)$ and the result follows.

$\square$

APPENDIX C. PROOF OF THEOREM 2

We use the same notation as in the proof of Theorem 1 with the addition of sub/superscripts indicating the appropriate subsample $k = a, b$, where $k^c = \{a, b\} \setminus \{k\}$.

Step 0.(Combining) In this step we combine both subsample estimators. Letting $\Upsilon^k = D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}D^k/n_k$, for $k = a, b$, so that we have

$$
\begin{aligned}
\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) &= ((n_a/n)\Upsilon^a + (n_b/n)\Upsilon^b)^{-1} \times \\
&\times ((n_a/n)\Upsilon^a\sqrt{n}(\check{\alpha}_a - \alpha_0) + (n_b/n)\Upsilon^b\sqrt{n}(\check{\alpha}_b - \alpha_0)) \\
&= (V'V/n + o_P(1))^{-1} \times \\
&\times ((n_a/n)\Upsilon^a\sqrt{n}(\check{\alpha}_a - \alpha_0) + (n_b/n)\Upsilon^b\sqrt{n}(\check{\alpha}_b - \alpha_0)) + o_P(1) \\
&= \{V'V/n\}^{-1} \times \{(1/\sqrt{2}) \times \mathbb{G}_{na}[v_i\zeta_i] + (1/\sqrt{2})\mathbb{G}_{nb}[v_i\zeta_i]\} + o_P(1) \\
&= \{V'V/n\}^{-1} \times \mathbb{G}_n[v_i\zeta_i] + o_P(1)
\end{aligned}
$$

where we are also using the fact that

$$
\mathbb{E}_{n_k}[\widehat{v}_i^2] - \mathbb{E}_{n_k}[v_i^2] = o_P(1), \quad k = a, b
$$

which follows similarly to the proofs given in Step 6.

Letting $\Gamma = \mathrm{diag}(\zeta_1^2, \ldots, \zeta_n^2)$, define

$$
Z_n = (\mathrm{E}[V'V/n]^{-1}\mathrm{E}[V'\Gamma V/n]\mathrm{E}[V'V/n]^{-1})^{-1/2}\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) = \mathbb{G}_n[z_{i,n}] + o_P(1),
$$

where $z_{i,n} = (\mathrm{E}[V'V/n]^{-1}\mathrm{E}[V'\Gamma V/n]\mathrm{E}[V'V/n]^{-1})^{-1/2}v_i\zeta_i/\sqrt{n}$ are i.n.i.d. with mean zero. We have that for some small enough $\delta > 0$

$$
\bar{\mathrm{E}}|z_{i,n}|^{2+\delta} \lesssim \bar{\mathrm{E}}\left[|v_i|^{2+\delta}|\zeta_i|^{2+\delta}\right] \lesssim \sqrt{\bar{\mathrm{E}}|v_i|^{4+2\delta}}\sqrt{\bar{\mathrm{E}}|\zeta_i|^{4+2\delta}} \lesssim 1,
$$

by Condition SM.

This condition verifies the Lyapunov condition and thus implies that $Z_n \to_d N(0, 1)$.

Step 1.(Main) For the subsample $k = a, b$ write $\check{\alpha}_k = \left[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}D^k/n_k\right]^{-1}\left[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}Y_1^k/n_k\right]$ so that

$$
\sqrt{n_k}(\check{\alpha}_k - \alpha_0) = \left[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}D^k/n_k\right]^{-1}[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}(g^k + \zeta^k)/\sqrt{n_k}] =: ii_k^{-1} \cdot i_k.
$$

By Steps 2 and 3, $ii_k = V^{k\prime}V^k/n_k + o_P(1)$ and $i_k = V^{k\prime}\zeta^k/\sqrt{n_k} + o_P(1)$. Next note that $V^{k\prime}V^k/n_k = \mathrm{E}[V^{k\prime}V^k/n_k] + o_P(1)$ by Chebyshev, and we have that $\bar{\mathrm{E}}_k[v_i^2\zeta_i^2]$ and $\mathrm{E}[V^{k\prime}V^k/n_k]$ are bounded from above and away from zero by assumption.

Step 2. (Behavior of i_k.) Decompose

$$
i_k = V^{k\prime}\zeta^k/\sqrt{n_k} + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}}_{=:i_{ka}} + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}\zeta^k/\sqrt{n_k}}_{=:i_{kb}} + \underbrace{V^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}}_{=:i_{kc}} - \underbrace{V^{k\prime}\mathcal{P}_{\widehat{I}^{k^c}}\zeta^k/\sqrt{n_k}}_{=:i_{kd}}.
$$

First, note that by Condition ASTESS(v) we have

$$|i_{ka}| = |m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}| = o_P(1).$$

Second, by the split sample construction, we have that $\widehat{I}^{k^c}$ is independent from ζ^k, and by assumption of the model m^k is also independent of ζ^k. Thus by Chebyshev inequality

$$|i_{kb}| \lesssim_P \|\mathcal{M}_{\widehat{I}^{k^c}}m^k/\sqrt{n_k}\| \lesssim_P \sqrt{s\log p/n} = o_P(1),$$

where the last relation follows by Step 4.

Third, using similar independence arguments, by Chebyshev and Step 5, conclude

$$|i_{kc}| \lesssim_P \|\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}\| \lesssim_P \sqrt{s\log p/n} = o_P(1).$$

Fourth, using that $\widehat{s} \lesssim_P s$ by Lemma 1 so that $\phi_{\min}^{-1}(\widehat{s}) \lesssim_P 1$ by condition SE, we have that

$$|i_{kd}| \leqslant |\tilde{\beta}_{V^k}(\widehat{I}^{k^c})\prime X^{k\prime}\zeta^k/\sqrt{n_k}| \lesssim_P \sqrt{s/n} = o_P(1),$$

by Chebyshev since $\|X^k\tilde{\beta}_{V^k}(\widehat{I}^{k^c})/\sqrt{n_k}\| \lesssim_P \sqrt{s/n_k}$ because of the independence of the two subsamples k and k^c.

Step 3. (Behavior of ii_k.) Since $ii_k = (m^k + V^k)\prime\mathcal{M}_{\widehat{I}^{k^c}}(m^k + V^k)/n_k$, decompose

$$ii_k = V^{k\prime}V^k/n_k + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}m^k/n_k}_{=:ii_{ka}} + \underbrace{2m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}V^k/n_k}_{=:ii_{kb}} - \underbrace{V^{k\prime}\mathcal{P}_{\widehat{I}^{k^c}}V^k/n_k}_{=:ii_{kc}}.$$

Then $|ii_{ka}| \lesssim_P [s\log(p \vee n)]/n_k = o_P(1)$ by Step 4, $|ii_{kb}| \lesssim_P [s\log(p \vee n)]/n_k = o_P(1)$ by reasoning similar to deriving the bound for $|i_{kb}|$, and $|ii_{kc}| \lesssim_P [s\log(p \vee n)]/n_k = o_P(1)$ by reasoning similar to deriving the bound for $|i_{kd}|$.

Step 4. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}^{k^c}}m^k\|$ and related quantities.) For $k = a, b$ note that

$$
\begin{aligned}
\|\mathcal{M}_{\widehat{I}^{k^c}}m^k\| &\leqslant \|\mathcal{M}_{\widehat{I}_1^{k^c}}m^k\| \\
&\leqslant \|X^k\tilde{\beta}_{D^{k^c}}(\widehat{I}_1^{k^c}) - m^k\| \\
&\leqslant \|X^k(\tilde{\beta}_{D^{k^c}}(\widehat{I}_1^{k^c}) - \beta_{m0})\| + \|R_m^k\|.
\end{aligned}
$$

since $\widehat{I}_1^{k^c} \subseteq \widehat{I}^{k^c}$. By Condition ASTESS(iii), we have $\|R_m^k/\sqrt{n_k}\| \lesssim \sqrt{s/n}$ and, also using Lemma 1 so that $\widehat{s}^k \lesssim_P s$, we have

$$
\begin{aligned}
\|X^k(\tilde{\beta}_{D^{k^c}}(\widehat{I}_1^{k^c}) - \beta_{m0})/\sqrt{n_k}\| &\leqslant \sqrt{\phi_{\max}(\widehat{s}^a)_k/\phi_{\min}(\widehat{s}^a)_{k^c}}\|X^{k^c}(\tilde{\beta}_{D^{k^c}}(\widehat{I}_1^{k^c}) - \beta_{m0})/\sqrt{n_k}\| \\
&\lesssim_P \sqrt{[s\log(p \vee n)]/n_k}
\end{aligned}
$$

since $\sqrt{\phi_{\max}(\widehat{s}^a)_k/\phi_{\min}(\widehat{s}^a)_{k^c}} \lesssim_P 1$ by condition SE. Thus we also have established

$$\|\tilde{\beta}_{m^{k^c}}(\widehat{I}^{k^c}) - \beta_{m0}\| \lesssim_P \sqrt{[s\log(p \vee n)]/n_k}.$$

Step 5. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}^{kc}} g^k\|$ and related quantities.) For $k = a, b$ note that

$$
\begin{aligned}
\|\mathcal{M}_{\widehat{I}^{kc}} g^k\| \;&\leqslant\; \|\mathcal{M}_{\widehat{I}_2^{kc}} g^k\| \\
&\leqslant\; \|X^k \tilde{\tilde{\beta}}_{Y_1^{kc}}(\widehat{I}_2^{kc}) - g^k\| \\
&\leqslant\; \|X^k(\tilde{\beta}_{Y_1^{kc}}(\widehat{I}_2^{kc}) - \beta_{g0})\| + \|R_g^k\| \\
&\lesssim_P\; \sqrt{s \log(p \vee n)}
\end{aligned}
$$

since $\widehat{I}_2^{kc} \subseteq \widehat{I}^{kc}$, the triangle inequality, $\|R_g^k/\sqrt{n_k}\| \lesssim \sqrt{s/n_k}$, and by Lemma 1, similarly to Step 4 using SE, it follows that $\|X^k(\tilde{\beta}_{Y_1^{kc}}(\widehat{I}_2^{kc}) - \beta_{g0})/\sqrt{n_k}\| \lesssim_P \sqrt{[s \log(p \vee n)]/n_k}$.

Step 6. (Variance Estimation.) Since $\widehat{s}^k \lesssim_P s = o(n)$, $(n_k - \widehat{s}^k - 1)/n_k = o_P(1)$, so we can use n as the denominator. Hence consider

$$
\begin{aligned}
\mathbb{E}_n[\widehat{v}_i^2] \;&=\; (n_a/n) D^{a\prime} \mathcal{M}_{\widehat{I}^b} D^a / n_a + (n_b/n) D^{b\prime} \mathcal{M}_{\widehat{I}^a} D^b / n_b \\
&=\; (n_a/n) ii_a + (n_b/n) ii_b = V'V/n + o_P(1) = \bar{\mathrm{E}}[v_i^2] + o_P(1)
\end{aligned}
$$

by Step 3 and $\bar{\mathrm{E}}[|v_i|^q] \lesssim 1$ for some $q > 4$ by condition SM.

Let $\tilde{v}_i = v_i + r_{mi}$ and $\tilde{\zeta}_i = \zeta_i + r_{gi}$. Recall that by Condition ASTESS(vi) we have $\mathbb{E}_{n_k}[\tilde{v}_i^2 \tilde{\zeta}_i^2] - \bar{\mathrm{E}}_k[v_i^2 \zeta_i^2] \to_P 0$ for subsample $k = a, b$. To show that $\mathbb{E}_{n_k}[\widehat{v}_i^2 \widehat{\zeta}_i^2] - \mathbb{E}_{n_k}[\tilde{v}_i^2 \tilde{\zeta}_i^2] \to_P 0$ we start applying a triangular inequality for each $k = a, b$

$$
|\mathbb{E}_{n_k}[\widehat{v}_i^2 \widehat{\zeta}_i^2 - \tilde{v}_i^2 \tilde{\zeta}_i^2]| \leqslant |\mathbb{E}_{n_k}[(\widehat{v}_i^2 - \tilde{v}_i^2)\tilde{\zeta}_i^2]| + |\mathbb{E}_{n_k}[\tilde{v}_i^2(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]| + |\mathbb{E}_{n_k}[(\widehat{v}_i^2 - \tilde{v}_i^2)(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]|.
$$

Then,

$$
\begin{aligned}
|\mathbb{E}_{n_k}[\tilde{v}_i^2(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]| \;&\leqslant\; 2\mathbb{E}_{n_k}[\{d_i(\alpha_0 - \check{\alpha}_k)\}^2 \tilde{v}_i^2] + 2\mathbb{E}_{n_k}[\{x_i'(\check{\beta}_k - \beta_{g0})\}^2 \tilde{v}_i^2] \\
&\quad + |2\mathbb{E}_{n_k}[\tilde{\zeta}_i d_i(\alpha_0 - \check{\alpha}_k)\tilde{v}_i^2]| + |2\mathbb{E}_{n_k}[\tilde{\zeta}_i x_i'(\check{\beta}_k - \beta_{g0})\tilde{v}_i^2]|.
\end{aligned}
$$

As a consequence of Condition SM we have $\mathrm{E}[\max_{i \leqslant n} d_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i \leqslant n} \zeta_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i \leqslant n} v_i^2] \lesssim n^{2/q}$, thus by Markov inequality we have $\|D\|_\infty + \|\zeta\|_\infty + \|V\|_\infty \lesssim_P n^{1/q}$. Therefor, by condition SM and ASTESS(vi) we have $(\|V\|_\infty^2 + \|R_m\|_\infty^2) s \log(p \vee n) = o_P(n)$. Thus we have the following relations:

$$
\begin{aligned}
\mathbb{E}_{n_k}[\{d_i(\alpha_0 - \check{\alpha}_k)\}^2 \tilde{v}_i^2] \;&\leqslant\; |\alpha_0 - \check{\alpha}_k|^2 \mathbb{E}_{n_k}[d_i^2] \max_{i \leqslant n} \tilde{v}_i^2 \lesssim_P n^{-1}(\|V\|_\infty^2 + \|R_m\|_\infty^2) = o_P(1), \\
\mathbb{E}_{n_k}[\{x_i'(\check{\beta}_k - \beta_{g0})\}^2 \tilde{v}_i^2] \;&\leqslant\; \max_{i \leqslant n} \tilde{v}_i^2 \mathbb{E}_{n_k}[\{x_i'(\check{\beta}_k - \beta_{g0})\}^2] \\
&\lesssim_P\; (\|V\|_\infty^2 + \|R_m\|_\infty^2)[s \log(p \vee n)]/n = o_P(1), \\
|\mathbb{E}_{n_k}[\tilde{\zeta}_i d_i(\alpha_0 - \check{\alpha}_k)\tilde{v}_i^2]| \;&\leqslant\; \max_{i \leqslant n} |\tilde{v}_i| \{\mathbb{E}_{n_k}[\{d_i(\alpha_0 - \check{\alpha}_k)\}^2] \mathbb{E}_{n_k}[\tilde{\zeta}_i^2 \tilde{v}_i^2]\}^{1/2} \\
&\lesssim_P\; (\|V\|_\infty + \|R_m\|)\sqrt{1/n} = o_P(1), \\
|\mathbb{E}_{n_k}[\tilde{\zeta}_i x_i'(\check{\beta}_k - \beta_{g0})\tilde{v}_i^2]| \;&\leqslant\; \max_{i \leqslant n} |\tilde{v}_i| \{\mathbb{E}_{n_k}[\{x_i'(\check{\beta}_k - \beta_{g0})\}^2] \mathbb{E}_{n_k}[\tilde{\zeta}_i^2 \tilde{v}_i^2]\}^{1/2} \\
&\lesssim_P\; (\|V\|_\infty + \|R_m\|_\infty)\sqrt{[s \log(p \vee n)]/n} = o_P(1),
\end{aligned}
$$

since $\mathbb{E}_{n_k}[\tilde{\zeta}_i^2 \tilde{v}_i^2] \lesssim_P 1$, $\|\check{\beta}_k - \beta_{g0}\|^2 \lesssim_P [s \log(p \vee n))]/n$ by Lemma 1, and $|\check{\alpha}_k - \alpha_0|^2 \lesssim_P 1/n$ by Step 1.

Similarly, $\mathbb{E}_{n_k}[(\widehat{v}_i^2 - v_i^2)\tilde{\zeta}_i^2] = o_P(1)$.

Finally, under condition ASTESS(vi) $\max_{i\leqslant n} \|(\widehat{v}_i, \widehat{\zeta}_i, \tilde{\zeta}_i, \tilde{v}_i)'\|_\infty s \log(p \vee n) = o_P(n)$

$$
\begin{aligned}
|\mathbb{E}_{n_k}[(\widehat{v}_i^2 - \tilde{v}_i^2)(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)]| \quad &\leqslant \{\mathbb{E}_{n_k}[(\widehat{v}_i^2 - \tilde{v}_i^2)^2]\mathbb{E}_{n_k}[(\widehat{\zeta}_i^2 - \tilde{\zeta}_i^2)^2]\}^{1/2} \\
&\leqslant \{\mathbb{E}_{n_k}[2(\widehat{v}_i^2 + \tilde{v}_i^2)(\widehat{v}_i - \tilde{v}_i)^2]\mathbb{E}_{n_k}[2(\widehat{\zeta}_i^2 + \tilde{\zeta}_i^2)(\widehat{\zeta}_i - \tilde{\zeta}_i)^2]\}^{1/2} \\
&\leqslant 4\max_{i\leqslant n} \|(\widehat{v}_i, \widehat{\zeta}_i, \tilde{\zeta}_i, \tilde{v}_i)'\|_\infty^2 \{\mathbb{E}_{n_k}[(\widehat{v}_i - \tilde{v}_i)^2]\mathbb{E}_{n_k}[(\widehat{\zeta}_i - \tilde{\zeta}_i)^2]\}^{1/2} \\
&\lesssim_P \max_{i\leqslant n} \|(\widehat{v}_i, \widehat{\zeta}_i, \tilde{\zeta}_i, \tilde{v}_i)'\|_\infty^2 [s \log(n \vee p)]/n = o_P(1).
\end{aligned}
$$

$\square$

References

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," arXiv:[math.ST], http://arxiv.org/abs/1012.1297.

Belloni, A., and V. Chernozhukov (2011a): "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models," *Annals of Statistics*, 39(1), 82–130.

———— (2011b): "High Dimensional Sparse Econometric Models: An Introduction," *Inverse problems and high dimensional estimation - Stats in the Château summer school in econometrics and statistics, 2009, Springer Lecture Notes in Statistics - Proceedings*, pp. 121–156.

———— (2011c): "Least Squares After Model Selection in High-dimensional Sparse Models," *forthcoming Bernoulli*.

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "LASSO Methods for Gaussian Instrumental Variables Models," arXiv:[math.ST], http://arxiv.org/abs/1012.1297.

———— (2011): "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics. 10th World Congress of Econometric Society.*

Belloni, A., V. Chernozhukov, and L. Wang (2010): "Square-Root-LASSO: Pivotal Recovery of Nonparametric Regression Functions via Conic Programming," *Duke and MIT Working Paper*.

———— (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37(4), 1705–1732.

Candes, E., and T. Tao (2007): "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Statist.*, 35(6), 2313–2351.

Card, D., and A. B. Krueger (1997): *Myth and Measurement: The New Economics of the Minimum Wage.* Princeton University Press.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

Donohue III, J. J., and S. D. Levitt (2001): "The Impact of Legalized Abortion on Crime," *Quarterly Journal of Economics*, 116(2), 379–420.

HECKMAN, J., R. LALONDE, AND J. SMITH (1999): "The economics and econometrics of active labor market programs," *Handbook of labor economics*, 3, 1865–2097.

IMBENS, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.

JING, B.-Y., Q.-M. SHAO, AND Q. WANG (2003): "Self-normalized Cramr-type large deviations for independent random variables," *Ann. Probab.*, 31(4), 2167–2215.

KOENKER, R. (1988): "Asymptotic Theory and Econometric Practice," *Journal of Aplpied Econometrics*, 3, 139–147.

MEINSHAUSEN, N., AND B. YU (2009): "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, 37(1), 2246–2270.

NEWEY, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

TIBSHIRANI, R. (1996): "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.