Review of Economic Studies (2014) **81**, 608–650 doi: 10.1093/restud/rdt044

Advance access publication 24 November 2013

Inference on Treatment Effects after Selection among High-Dimensional Controls[†]

ALEXANDRE BELLONI

Duke University

VICTOR CHERNOZHUKOV

MIT

and

CHRISTIAN HANSEN

University of Chicago

First version received October 2012*; final version accepted October* 2013 (*Eds.*)

We propose robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of very many regressors in a model with possibly non-Gaussian and heteroscedastic disturbances. We allow for the number of regressors to be larger than the sample size. To make informative inference feasible, we require the model to be approximately sparse; that is, we require that the effect of confounding factors can be controlled for up to a small approximation error by including a relatively small number of variables whose identities are unknown. The latter condition makes it possible to estimate the treatment effect by selecting approximately the right set of regressors. We develop a novel estimation and uniformly valid inference method for the treatment effect in this setting, called the "post-double-selection" method. The main attractive feature of our method is that it allows for imperfect selection of the controls and provides confidence intervals that are valid uniformly across a large class of models. In contrast, standard post-model selection estimators fail to provide uniform inference even in simple cases with a small, fixed number of controls. Thus, our method resolves the problem of uniform inference after model selection for a large, interesting class of models. We also present a generalization of our method to a fully heterogeneous model with a binary treatment variable. We illustrate the use of the developed methods with numerical simulations and an application that considers the effect of abortion on crime rates.

Key words: Treatment effects, Partially linear model, High-dimensional-sparse regression, Inference under imperfect model selection, Uniformly valid inference after model selection, Average treatment effects, Lasso, Orthogonality of estimating equations with respect to nuisance parameters.

JEL Codes: C01

1. INTRODUCTION

Many empirical analyses focus on estimating the structural, causal, or treatment effect of some variable on an outcome of interest. For example, we might be interested in estimating the causal

[†] This is a revision of a 2011 ArXiv/CEMMAP paper entitled "Estimation of Treatment Effects with High-Dimensional Controls".

effect of some government policy on an economic outcome such as 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 important method is based on the assumption that the variable of interest can be taken as randomly assigned after controlling for a sufficient set of other factors; see, for example, Heckman *et al*. (1999) and Imbens (2004).

A problem empirical researchers face when relying on a conditional-on-observables identification strategy for estimating a structural effect is knowing which controls to include. 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 selecting a set of controls from a potentially vast set of 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 an *ad hoc* 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 article, or examples in Angrist and Pischke (2008) among many other references.

We present an approach to estimating and performing inference on structural effects in an environment where 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

$$y_i = d_i\alpha_0 + g(z_i) + \zeta_i \tag{1.1}$$

where d_i is the treatment/policy variable of interest, z_i is a set of control variables, and ζ_i is an unobservable that satisfies $\mathrm{E}[\zeta_i \,|\, d_i, z_i] = 0$.[1] The goal of the econometric analysis is to conduct inference on the treatment effect α_0. We examine the problem of selecting a set of variables from among p potential regressors $x_i = P(z_i)$, which may consist of z_i and transformations of z_i, to adequately approximate $g(z_i)$ allowing for $p > n$. Of course, useful inference about α_0 is unavailable in this framework without imposing further structure. We impose such structure by assuming that exogeneity of d_i may be taken as given once one controls linearly for a relatively small number $s < n$ of variables in x_i whose identities are *a priori* unknown. This assumption implies that linear combinations of these s unknown regressors provide approximations to $g(z_i)$ and to $\mathrm{E}[d_i|z_i] = m(z_i)$ which produce relatively small approximation errors for each object. This assumption, which is termed approximate sparsity or simply sparsity, allows us to approach the problem of estimating α_0 as a variable selection problem. This framework includes as special cases the most common approaches to parametric and nonparametric regression analysis and allows for the realistic scenario in which the researcher is unsure about exactly which variables or transformations are important confounds and so must search among a broad set of controls.[2]

[1] We note that d_i does not need to be binary. This structure may also arise in the context of randomized treatment in the case where treatment assignment depends on underlying control variables, potentially in a complicated way. See, for example, Duflo *et al*. (2008), especially Section 6.1, and Kremer and Glennerster (2011).

[2] High-dimensional x_i typically occurs in either of two ways. First, the baseline set of conditioning variables itself may be large so $x_i = z_i$, and we assume $g(z_i) = g(x_i) \approx x_i'\beta_g$. Second, z_i may be low-dimensional, but one may wish to entertain many non-linear transformations of z_i in forming $x_i = P(z_i)$ as in traditional series-based estimation of the partially linear model. In the second case, one might prefer to refer to z_i as the controls and x_i as something else, such as technical regressors. For simplicity of exposition and as the formal development in the article is agnostic about the source of high-dimensional x_i, we call the variables in x_i controls or control variables in either case.

The main contributions of this article are providing an estimation and inference method within a partially linear model with potentially very high-dimensional controls and developing the supporting theory establishing its validity uniformly across a rich class of data-generating processes (dgps). Our approach differs from usual post-model-selection methods that rely on a single selection step. Rather, we use two different variable selection steps followed by a final estimation step as follows:

1. In the first step, we select a set of control variables that are useful for predicting the treatment d_i. This step helps to insure validity of post-model-selection-inference by finding control variables that are strongly related to the treatment and thus potentially important confounding factors.
2. In the second step, we select additional variables by selecting control variables that predict y_i. 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 providing an additional chance to find important confounds.
3. In the final step, we estimate the treatment effect α_0 of interest by the linear regression of y_i on the treatment d_i and the union of the set of variables selected in the two variable selection steps.

We provide theoretical results on the properties of the resulting treatment effect estimator and show that it provides inference that is uniformly valid over large classes of models and also achieves 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 heteroscedasticity of the model's errors.

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 original paper. By using formal variable selection, we select a small set of between eight and twelve variables depending on the outcome, compared to the set of eight variables considered by Donohue III and Levitt (2001). Once this set of variables is linearly controlled for, the estimated abortion effect is rendered imprecise. The selected variables differ substantially from the eight variables used in Donohue III and Levitt (2001) and are generally related to non-linear trends that depend on initial state-level characteristics. It is interesting that Foote and Goetz (2008) raise a similar point based on intuitive grounds and additional data in a comment on Donohue III and Levitt (2001). Foote and Goetz (2008) find that a linear trend interacted with crime rates computed before abortion could have had an effect renders the estimated abortion effects imprecise.[3] Overall, finding that a formal, rigorous approach to variable selection produces a qualitatively different result than a more *ad hoc* approach suggests 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.

[3] Donohue III and Levitt (2008) provide yet more data and a more flexible specification in response to Foote and Goetz (2008). In a supplement available at http://faculty.chicagobooth.edu/christian.hansen/research/, we provide additional results based on Donohue III and Levitt (2008). The conclusions are similar to those obtained in this article in that we find the estimated abortion effect becomes imprecise once one allows for a broad set of controls and selects among them. However, the specification of Donohue III and Levitt (2008) relies on a large number of district cross time fixed effects and so does not immediately fit into our regularity conditions. We conjecture the methodology continues to work in this case but leave verification to future research.

Relationship to literature. We contribute to several existing literatures. First, we contribute to the literature on semi-parametric estimation of partially linear models; see Donald and Newey (1994), Härdle *et al*. (2000), Robinson (1988), and others.[4] We differ from most of the existing literature that considers $p \ll n$ series terms by allowing $p \gg n$ series terms from which we select $\widehat{s} \ll n$ terms to construct the regression fits. Considering an initial broad set of terms allows for more refined approximations of regression functions relative to the usual approach that uses only a few low-order terms. See, for example, Belloni *et al*. (2011) for a wage function example and Section 4 for theoretical examples. However, our most important contribution is to allow for data-dependent selection of the appropriate series terms. The previous literature on inference in the partially linear model generally takes the series terms as given without allowing for their data-driven selection. However, selection of series terms is crucial for achieving consistency when $p \gg n$ and is needed for increasing efficiency even when $p = Cn$ with $C < 1$.[5]

Second, we contribute to the literature on the estimation of treatment effects. We note that the policy variable d_i does not have to be binary in our framework. However, our method has a useful interpretation related to the propensity score when d_i is binary. In the first selection step, we select terms from x_i that predict the treatment d_i, *i.e.* terms that explain the propensity score. We also select terms from x_i that predict y_i, *i.e.* terms that explain the outcome regression function. Then we run a final regression of y_i on the treatment d_i and the union of selected terms. Thus, our procedure relies on the selection of variables relevant for both the propensity score and the outcome regression and is related to treatment effects estimators that use regression adjustment after conditioning on the propensity score. Relying on selecting variables that are important for both objects allows us to achieve two goals: we obtain uniformly valid confidence sets for α_0 despite imperfect model selection, and we achieve full efficiency for estimating α_0 in the homoscedastic case. The relation of our approach to the propensity score brings about interesting connections to the treatment effects literature. Hahn (1998), Heckman *et al*. (1998), and Abadie and Imbens (2011) have constructed efficient regression or matching-based estimates of average treatment effects. Hahn (1998) also shows that conditioning on the propensity score is unnecessary for efficient estimation of average treatment effects. Hirano *et al*. (2003) demonstrate that one can efficiently estimate average treatment effects using estimated propensity score weighting alone. Robins and Rotnitzky (1995) have shown that using propensity score modeling coupled with a parametric regression model leads to efficient estimates if either the propensity score model or the parametric regression model is correct. While our contribution is quite distinct from these approaches, it also highlights the important robustness role played by the propensity score model in the selection of the right control terms for the final regression.

Third, we contribute to the literature on estimation and inference with high-dimensional data and to the uniformity literature. There has been extensive work on estimation and perfect model selection in both low and high-dimensional contexts; see, *e.g*. Hansen (2005) and Belloni *et al*. (2010) for reviews focused on econometric applications. However, there has been little work on inference after imperfect model selection. Perfect model selection relies on extremely unrealistic assumptions, and even moderate model selection mistakes can have serious consequences for inference as has been shown in Pötscher (2009), Leeb and Pötscher (2008a), and others. In work on

[4] Following Robinson (1988)'s method, estimation of the parameters of the linear part of a partially linear model is typically done by regressing $y_i - \widehat{\mathrm{E}}[y_i|z_i]$ on $d_i - \widehat{\mathrm{E}}[d_i|z_i]$ where $\widehat{\mathrm{E}}[y_i|z_i]$ and $\widehat{\mathrm{E}}[d_i|z_i]$ are preliminary non-parametric estimators of the conditional expectations of y_i and d_i given z_i under the assumption that $\dim(z_i)$ is small. Our approach implicitly fits within this framework where we are offering selection based estimators of the conditional expectation functions.

[5] Cattaneo *et al*. (2010) derive properties of series estimator under $p = Cn$, $C < 1$, asymptotics. It follows from their results that, under homoscedasticity, the series estimator achieves the semiparametric efficiency bound only if $C \to 0$.

instrument selection for estimation of a linear instrumental variables model, Belloni *et al*. (2010) and Belloni *et al*. (2012) have shown that moderate model selection mistakes do not prevent valid inference about low-dimensional structural parameters by exploiting the orthogonality or "immunization" property of the problem, whereby the moment equation identifying the target parameter is not affected by small perturbations of the nuisance function, the optimal instrument in the IV context.[6] The partially linear regression model (1.1) does not immediately have the same orthogonality structure, and model selection based on the outcome regression alone produces confidence intervals with poor coverage properties. However, our post-double selection procedure, which also selects controls that explain $\mathrm{E}[d_i|z_i]$, creates the necessary orthogonality by performing two separate model selection steps. Performing the two selection steps helps reduce omitted variable bias so that it is possible to perform uniform inference after model selection.[7] In that regard, our contribution is in the spirit of and builds upon the classical contribution by Romano (2004) on the uniform validity of t-tests for the univariate mean. It also shares the spirit of recent contributions, among others, by Mikusheva (2007) on uniform inference in autoregressive models, by Andrews and Cheng (2011) on uniform inference in moment condition models that are potentially unidentified, and by Andrews *et al*. (2011) on a generic framework for uniformity analysis.

Finally, we contribute to the broader literature on high-dimensional estimation. For variable selection, we use ℓ_1-penalized methods, though our method and theory will allow for the use of other methods. ℓ_1-penalized methods have been proposed for model selection problems in high-dimensional least squares problems, *e.g*. Lasso in Frank and Friedman (1993) and Tibshirani (1996), in part because they are computationally efficient. Many ℓ_1-penalized methods and related methods have been shown to have good estimation properties even when perfect variable selection is not feasible; see, *e.g*. Candès and Tao (2007), Meinshausen and Yu (2009), Bickel *et al*. (2009), Huang *et al*. (2010), Belloni and Chernozhukov (2013) and the references therein. Such methods have also been shown to extend to non-parametric and non-Gaussian cases as in Bickel *et al*. (2009) and Belloni *et al*. (2012). 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, it corresponds to the usual approach in economics that relies on considering a small number of controls.

Notation. We work with triangular array data $\{\omega_{i,n}\}_{i=1}^{n}$, which is an observable set of the first n elements of the infinite data stream $\{\omega_{i,n}\}_{i=1}^{\infty}$ defined on the infinite product probability space $(\Omega, \mathcal{A}, \mathrm{P}_n)$, where $\mathrm{P}=\mathrm{P}_n$ the probability measure or data-generating process for the entire infinite stream can change with n. We shall use $\mathbf{P}_n$ (possibly dependent on n) as the sets of potential probability measures P that satisfy certain assumptions. Each $\omega_{i,n}=(y'_{i,n}, z'_{i,n}, d'_{i,n})'$ is a vector with components defined below, and these vectors are i.n.i.d.—independent across i, but not necessarily identically distributed. Thus, all parameters that characterize the distribution of $\{\omega_{i,n}\}_{i=1}^{\infty}$ are implicitly indexed by P_n and thus by the sample size n. We shall omit the dependence on n and on P_n from the notation where possible. We use such array asymptotics as doing so allows consideration of approximating sequences that better capture some finite-sample phenomena and to insure the robustness of conclusions with respect to perturbations of the data-generating process P along various sequences. This robustness, in turn, translates into uniform validity of confidence regions over certain regions $\mathbf{P}=\cap_{n\geqslant n_0}\mathbf{P}$ of data-generating processes, where $n_0\geqslant 1$ is a fixed sample size.

[6] To the best of our knowledge, Belloni *et al*. (2010) and Belloni *et al*. (2012) were the first to use this immunization/orthoganility property in the $p \gg n$ setup. We provide a further discussion on this property in Section 5.

[7] Note that this claim only applies to the main parameter α_0 and does not apply to the nuisance part g. Furthermore, our claim of uniformity only applies to models in which both g and m are approximately sparse. See the remarks following Theorem 1 for further discussion.

We 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, denoted $\|\cdot\|_0$, is 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 $j \in T$ and $\delta_{Tj} = 0$ 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. We also use $\rightsquigarrow$ to denote convergence in distribution. Given a p-vector b, we denote support$(b) = \{j \in \{1,...,p\} : b_j \neq 0\}$.

2. INFERENCE ON TREATMENT AND STRUCTURAL EFFECTS CONDITIONAL ON OBSERVABLES

2.1. *Framework*

We consider the partially linear model

$$y_i = d_i\alpha_0 + g(z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \,|\, z_i, d_i] = 0, \tag{2.2}$$

$$d_i = m(z_i) + v_i, \quad \mathrm{E}[v_i \,|\, z_i] = 0, \tag{2.3}$$

where y_i is the outcome variable, d_i is the policy/treatment variable whose impact α_0 we would like to infer,[8] z_i represents confounding factors on which we need to condition, and ζ_i and v_i are disturbances.

The confounding factors z_i affect the policy variable via the function $m(z_i)$ and the outcome variable via the function $g(z_i)$. Both of these functions are unknown and potentially complicated. We use linear combinations of control terms $x_i = P(z_i)$ to approximate $g(z_i)$ and $m(z_i)$, writing (2.2) and (2.3) as

$$y_i = d_i\alpha_0 + \underbrace{x_i'\beta_{g0} + r_{gi}}_{g(z_i)} + \zeta_i, \tag{2.4}$$

$$d_i = \underbrace{x_i'\beta_{m0} + r_{mi}}_{m(z_i)} + v_i, \tag{2.5}$$

where $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ are 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 pertinent confounding factors, the vector of controls, $x_i = P(z_i)$, can have dimension $p = p_n$ which can be large relative to the sample size. Specifically, our results require $\log p = o(n^{1/3})$ along with other technical conditions. High-dimensional regressors $x_i = P(z_i)$ could arise for different reasons. For instance, the list of available controls could be large, *i.e.* $x_i = z_i$ as in *e.g.* Koenker (1988). It could also be that many technical controls are present; *i.e.* the list $x_i = P(z_i)$ could be composed of a large number of transformations of elementary regressors z_i such as B-splines, dummies, polynomials, and various interactions as in Newey (1997), Chen (2007), or Chen and Pouzo (2009; 2012).

[8] We consider the case where d_i is a scalar for simplicity. Extension to the case where d_i is a vector of fixed, finite dimension is accomplished by introducing an equation like (2.3) for each element of the vector.

Having very many controls creates a challenge for estimation and inference. A key condition that makes it possible to perform constructive estimation and inference in such cases is termed sparsity. Sparsity is the condition that there exist approximations $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ to $g(z_i)$ and $m(z_i)$ in (2.4)–(2.5) that require only a small number of non-zero coefficients to render the approximation errors r_{gi} and r_{mi} small relative to estimation error. More formally, sparsity relies on two conditions. First, 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 so that

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

Second, the sparsity condition requires the size of the resulting approximation errors to be small compared to the conjectured size of the estimation error:

$$\{\bar{\mathrm{E}}[r_{gi}^2]\}^{1/2} \lesssim \sqrt{s/n} \text{ and } \{\bar{\mathrm{E}}[r_{mi}^2]\}^{1/2} \lesssim \sqrt{s/n}.$$

Note that the size of the approximating model $s = s_n$ can grow with n just as in standard series estimation.

The high-dimensional-sparse-model framework outlined above extends the standard framework in the treatment effect literature which assumes both that the identities of the relevant controls are known and that the number of such controls s is much smaller than the sample size. Instead, we assume that there are many, p, potential controls of which at most s controls suffice to achieve a desirable approximation to the unknown functions $g(\cdot)$ and $m(\cdot)$ and allow the identity of these controls to be unknown. Relying on this assumed sparsity, we use selection methods to select approximately the right set of controls and then estimate the treatment effect α_0.

2.2. *The method: least squares after double selection*

To define the method, we first write the reduced form corresponding to (2.2)–(2.3) as

$$y_i = x_i'\bar{\beta}_0 + \bar{r}_i + \bar{\zeta}_i, \tag{2.6}$$

$$d_i = x_i'\beta_{m0} + r_{mi} + v_i, \tag{2.7}$$

where $\bar{\beta}_0 := \alpha_0\beta_{m0} + \beta_{g0}$, $\bar{r}_i := \alpha_0 r_{mi} + r_{gi}$, $\bar{\zeta}_i := \alpha_0 v_i + \zeta_i$. We have two equations and hence can apply model selection methods to each equation to select control terms. Given the set of selected controls from (2.6) and (2.7), we can estimate α_0 by a least squares regression of y_i on d_i and the union of the selected controls. Inference on α_0 may then be performed using conventional methods for inference about parameters estimated by least squares.

The most important 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. Intuitively, this procedure works well since we are more likely to recover key controls by considering selection of controls from both equations instead of just considering selection of controls from the single equation (2.4) or (2.6). In finite-sample experiments, single-selection methods essentially fail, providing poor inference relative to the double-selection method outlined above. This performance is also supported theoretically by the fact that the double-selection method requires weaker regularity conditions for its validity and for attaining the semi-parametric efficiency bound[9] than the single selection method.

[9] The semi-parametric efficiency bound of Robinson (1988) is attained in the homoscedastic case whenever such a bound formally applies.

Now we formally define the post-double-selection estimator: Let $\widehat{I}_1$ denote the control terms selected by a variable selector computed using data $(\tilde{y}_i, \tilde{x}_i) = (d_i, x_i)$, $i = 1, ..., n$, and let $\widehat{I}_2$ denote the control terms selected by a variable selector computed using data $(\tilde{y}_i, \tilde{x}_i) = (y_i, x_i)$, $i = 1, ..., n$. The post-double-selection estimator $\check{\alpha}$ of α_0 is defined as the least squares estimator obtained by regressing y_i on d_i and the selected control terms x_{ij} with $j \in \widehat{I} \supseteq \widehat{I}_1 \cup \widehat{I}_2$:

$$(\check{\alpha}, \check{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\operatorname{argmin}} \{\mathbb{E}_n[(y_i - d_i\alpha - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}. \tag{2.8}$$

The set $\widehat{I}$ may contain variables that were not selected in the variable selection steps with indices in a set, say $\widehat{I}_3$, that the analyst thinks are important for ensuring robustness. We call $\widehat{I}_3$ the amelioration set. Thus,

$$\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2 \cup \widehat{I}_3; \tag{2.9}$$

let $\widehat{s} = \|\widehat{I}\|_0$ and $\widehat{s}_j = \|\widehat{I}_j\|_0$ for $j = 1, 2, 3$. We define a feasible Lasso estimator below and focus on the use of feasible Lasso for variable selection in the majority of results in this article. When feasible Lasso is used to construct $\widehat{I}_1$ and $\widehat{I}_2$, we refer to the post-double-selection estimator as the post-double-Lasso estimator. When other model selection devices are used to construct $\widehat{I}_1$ and $\widehat{I}_2$, we refer to the estimator as the generic post-double-selection estimator.

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

$$([\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1) \tag{2.10}$$

under approximate sparsity conditions, uniformly within a rich set of data-generating processes. We also show that the standard plug-in estimator for standard errors is consistent in these settings. Figure 1 (right panel) illustrates the result (2.10) by showing that the finite-sample distribution of our post-double-Lasso estimator is very close to the normal distribution. In contrast, Figure 1 (left panel) illustrates the problem with the traditional post-single-selection estimator based on (2.4), showing that its distribution is bimodal and sharply deviates from the normal distribution.

2.3. *Selection of controls via feasible Lasso methods*

Here we describe feasible variable 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. We use the version of the Lasso estimator from Belloni *et al.* (2012) geared for heteroscedastic, non-Gaussian cases, which solves

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

where $\widehat{\Psi} = \mathrm{diag}(\widehat{l}_1, \ldots, \widehat{l}_p)$ is a diagonal matrix of penalty loadings and $\|\widehat{\Psi}\beta\|_1 = \sum_{j=1}^p |\widehat{l}_j\beta_j|$. The penalty level λ and loadings $\widehat{l}_j$'s are set as

$$\lambda = 2 \cdot c\sqrt{n}\Phi^{-1}(1 - \gamma/2p) \text{ and } \widehat{l}_j = l_j + o_P(1), \;\; l_j = \sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2\epsilon_i^2]}, \text{ uniformly in } j = 1, \ldots, p, \tag{2.12}$$

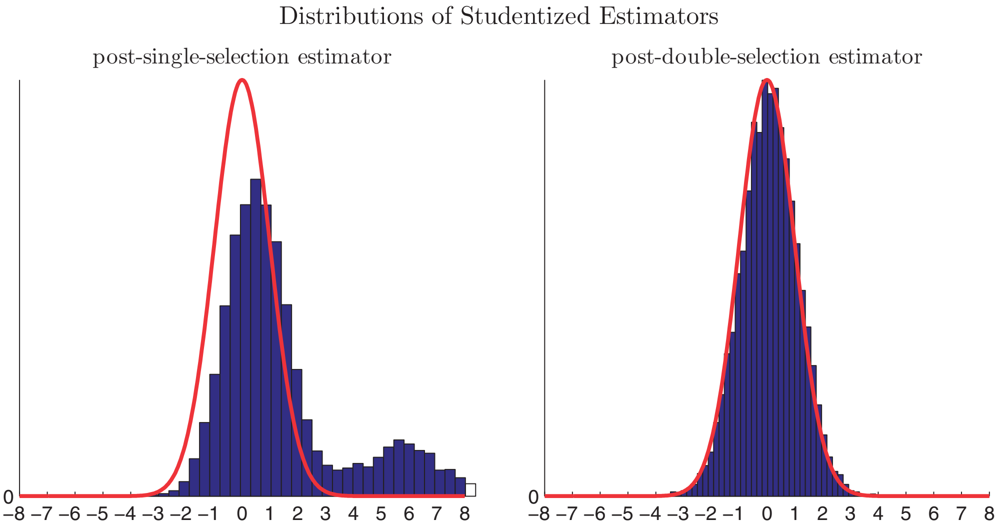

FIGURE 1

The finite-sample distributions (densities) of the standard post-single selection estimator (left panel) and of our proposed post-double selection estimator (right panel). The distributions are given for centered and studentized quantities. The results are based on 10000 replications of Design 1 described in Section 4.2, with R^2's in equation (2.6) and (2.7) set to 0.5.

where $c>1$ and $1-\gamma$ is a confidence level.[10] The l_j's are ideal penalty loadings that are not observed, and we estimate l_j by $\widehat{l_j}$ obtained via an iteration method given in Appendix A. The validity of using these estimates was established in Belloni *et al*. (2012) Lemma 11.[11]

A feature of the Lasso estimator is that the non-differentiability of the penalty function at zero induces the solution $\widehat{\beta}$ to have components set exactly to zero, and thus the Lasso solution may be used for model selection. In this article, we use $\widehat{\beta}$ as a model selection device. Specifically, we only make use of

$$\widehat{T} = \text{support}(\widehat{\beta}),$$

the labels of the regressors with non-zero estimated coefficients. We show that the selected model $\widehat{T}$ has good approximation properties for the regression function f under approximate sparsity in Section 3.[12] In what follows, we use the term feasible Lasso to refer to a Lasso estimator $\widehat{\beta}$ solving (2.11)–(2.12) with $c>1$ and $1-\gamma$ set such that

$$\gamma = o(1) \text{ and } \log(1/\gamma) \lesssim \log(p \vee n). \tag{2.13}$$

[10] Practical recommendations include the choice $c=1.1$ and γ close to zero, for example $\gamma=(1/n)\wedge .05$.

[11] Other methods that provably can be used in the heteroscedastic, non-Gaussian cases in the present context are the square-root-Lasso estimator (Belloni *et al*., 2011) and self-tuned Dantzig estimator (Gautier and Tsybakov, 2011).

[12] In estimating the l_j's and in Section 5, we also make use of the post-Lasso or Gauss-Lasso estimator as in Belloni and Chernozhukov (2013) which is obtained by running conventional least squares of the outcome on just the variables that were estimated to have non-zero coefficients by Lasso and using zero for the rest of the coefficients.

2.4. *Intuition for the importance of double selection*

To build intuition, we discuss the issues surrounding post-model selection inference in the case where there is only one control. In this simple example, we illustrate that a key defect of single-selection methods is that they fail to control omitted variables bias, and we demonstrate how double-selection helps overcome this problem. With one control, the model is

$$y_i = \alpha_0 d_i + \beta_g x_i + \zeta_i, \tag{2.14}$$

$$d_i = \beta_m x_i + v_i. \tag{2.15}$$

For simplicity, all errors and controls are taken as normal,

$$\begin{pmatrix} \zeta_i \\ v_i \end{pmatrix} | x_i \sim N\left(0, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix}\right), \quad x_i \sim N(0,1), \tag{2.16}$$

where the variance of x_i is normalized to be 1. The underlying probability space is equipped with probability measure P. Let **P** denote the collection of all dgps P where (2.14)–(2.16) hold with non-singular covariance matrices in (2.16). Suppose that we have an i.i.d. sample $\{(y_i, d_i, x_i)\}_{i=1}^n$ obeying the dgp $\mathrm{P}_n \in \mathbf{P}$. The subscript n signifies that the dgp and all true parameter values may change with n to better model finite-sample phenomena such as coefficients being “close to zero”. As in the rest of the article, we keep the dependence of the true parameter values on n implicit. Under the stated assumption, x_i and d_i are jointly normal with variances $\sigma_x^2 = 1$ and $\sigma_d^2 = \beta_m^2 \sigma_x^2 + \sigma_v^2$ and correlation $\rho = \beta_m \sigma_x / \sigma_d$.

The standard post-single-selection method for inference proceeds by applying a model selection method to equation (2.14) only, followed by applying OLS to the selected model. In the model selection stage, standard selection methods would omit x_i wp $\to 1$ if

$$|\beta_g| \leqslant \frac{\ell_n}{\sqrt{n}} c_n, \; c_n := \frac{\sigma_\zeta}{\sigma_x \sqrt{1-\rho^2}}, \text{ for some } \ell_n \to \infty, \tag{2.17}$$

where ℓ_n is a slowly varying sequence depending only on **P**. On the other hand, these methods would include x_i wp $\to 1$ if

$$|\beta_g| \geqslant \frac{\ell_n'}{\sqrt{n}} c_n, \text{ for some } \ell_n' > \ell_n, \tag{2.18}$$

where ℓ_n' is another slowly varying sequence in n depending only on **P**. As an example, one could do model selection in the $p=1$ case with a conservative t-test which drops x_i if the t-statistic $|t| = |\widehat{\beta}_g| / \text{s.e.}(\widehat{\beta}_g) \leqslant \Phi^{-1}(1-\gamma/2)$ where $\gamma = 1/n$, $\widehat{\beta}_g$ is the OLS estimator, and s.e.$(\widehat{\beta}_g)$ is the conventional OLS standard error estimator.[13] With this choice of γ, we have $\Phi^{-1}(1-\gamma/2) = \sqrt{2\log n}(1+o(1))$, and we could then take $\ell_n = \sqrt{\log n}$ and $\ell_n' = 2\sqrt{\log n}$. Note that Lasso selectors which we employ in our formal analysis act much like conservative t-tests with critical value $\sqrt{2\log n}(1+o(1))$ in low-dimensional settings, so our discussion here applies if Lasso selection is used in place of the conservative t-test.

The behaviour of the resulting post-single-selection estimator, $\widehat{\alpha}$ is then heavily influenced by the sequence of underlying dgps. Under sequences of models P_n such that (2.18) holds x_i is

[13] Such a t-test is conservative in the sense that the false rejection probability is tending to zero.

included wp → 1. The post-single-selection estimator is then the OLS estimator including both d_i and x_i and follows standard large sample asymptotics under P_n:

$$\sigma_n^{-1}\sqrt{n}(\widehat{\alpha}-\alpha_0)=\underbrace{\sigma_n^{-1}\mathbb{E}_n[v_i^2]^{-1}\sqrt{n}\mathbb{E}_n[v_i\zeta_i]}_{=:i}+o_P(1)\rightsquigarrow N(0,1)$$

where $\sigma_n^2=\sigma_\zeta^2(\sigma_v^2)^{-1}$ is the semi-parametric efficiency bound for estimating α_0 under homoscedasticity. When $\beta_g=o(1/\sqrt{n})$ and ρ is bounded away from 1, x_i is excluded wp → 1. In this case, β_g is small enough that failing to control for x_i does not introduce large omitted variables bias, and the estimator satisfies

$$\sigma_n^{*-1}\sqrt{n}(\widehat{\alpha}-\alpha_0)=\underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i\zeta_i]}_{:=i^*}+o_P(1)\rightsquigarrow N(0,1)$$

where $\sigma_n^{*2}=\sigma_\zeta^2(\sigma_d^2)^{-1}\leq\sigma_n^2$. That is, the post-single-selection estimator may achieve a variance smaller than the semi-parametric efficiency bound under such coefficient sequences. The potential reduction in variance is often used to motivate single-selection procedures.

This "too good" behaviour of the single-selection procedure has its price, as emphasized in Leeb and Pötscher (2008b): There are plausible sequences of dgps P_n where the post-single-selection estimator $\widehat{\alpha}$ performs very poorly. For example, consider $\beta_g=\frac{\ell_n}{\sqrt{n}}c_n$, so the coefficient on the control is "moderately close to zero". In this case, the t-test set-up above cannot distinguish this coefficient from 0, and the control x_i is dropped wp → 1. It follows that[14]

$$|\sigma_n^{*-1}\sqrt{n}(\widehat{\alpha}-\alpha_0)|\rightsquigarrow\infty. \tag{2.19}$$

This poor behaviour occurs because the omitted variable bias created by dropping x_i scaled by $\sqrt{n}$ diverges to infinity, namely $|\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_ix_i]\beta_g|\propto\ell_n\to\infty$. That is, the standard post-selection estimator is not asymptotically normal and even fails to be consistent at the rate of $\sqrt{n}$ under this sequence and many other sequences with small but non-zero β_g. A similar argument can be used to show a similar failure of single-selection based solely on (2.15).

The post-double-selection estimator, $\check{\alpha}$ resolves this problem by doing variable selection via standard t-tests or Lasso-type selectors with two equations that contain the information from (2.14) and (2.15) and then estimating α_0 by regressing y_i on d_i and the union of the selected controls. By doing so, x_i is omitted only if its coefficient in both equations is small which greatly limits the potential for omitted variables bias. Formally, we drop x_i with positive probability only if

$$\text{both } |\beta_g|<\frac{\ell_n'}{\sqrt{n}}c_n \text{ and } |\beta_m|<\frac{\ell_n'}{\sqrt{n}}(\sigma_v/\sigma_x). \tag{2.20}$$

Given this property, it follows that the post-double selection estimator satisfies

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0)=i+o_P(1)\rightsquigarrow N(0,1), \tag{2.21}$$

under any sequence of $\mathrm{P}_n\in\mathbf{P}$ implying we get the same approximating distribution whether or not x_i is omitted. That $\check{\alpha}$ follows (2.21) when x_i is included is obvious as in the single-selection

[14] Indeed, note that $\sigma_n^{*-1}\sqrt{n}(\widehat{\alpha}-\alpha_0)=\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i\zeta_i]+\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_ix_i]\beta_g:=i^*+ii^*$. The term i^* has standard behaviour; namely $i^*\rightsquigarrow N(0,1)$. The term ii^* generates omitted variable bias, and it may be arbitrarily large since, wp → 1, $|ii^*|\geqslant\frac{1}{2}\frac{|\rho|}{\sqrt{1-\rho^2}}\ell_n\nearrow\infty$, if $\ell_n|\rho|\nearrow\infty$.

case. When x_i is dropped, we have

$$\sigma_n^{*-1}\sqrt{n}(\check{\alpha}-\alpha_0)=\underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i\zeta_i]}_{=i^*}+\underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_ix_i]\beta_g}_{=ii}.$$

Term *ii* arises due to omitted variables bias and leads to the divergent behaviour of the single-selection estimator. However, for the double-selection-estimator, we know that (2.20) holds when x_i is omitted; so we have wp → 1

$$|ii|\leqslant 2\sigma_\zeta^{-1}\sigma_d\sigma_d^{-2}\sqrt{n}\sigma_x^2|\beta_m\beta_g|\leqslant 2\frac{\sigma_v/\sigma_d}{\sqrt{1-\rho^2}}\frac{\ell_n'^{\,2}}{\sqrt{n}}=2\frac{(\ell_n')^2}{\sqrt{n}}\to 0.$$

for sensible ℓ_n' such as $\ell_n'\propto\sqrt{\log n}$ as above. Moreover, we can show $i^*-i=o_P(1)$ under such sequences, so the first-order asymptotics of $\check{\alpha}$ is the same whether x_i is included or excluded.

To summarize, the post-single-selection estimator may not be root-n consistent in sensible models which translates into bad finite-sample properties. The potential poor finite-sample performance may be clearly seen in Monte-Carlo experiments. The estimator $\widehat{\alpha}$ is thus non-regular: its first-order asymptotic properties depend on the model sequence P_n in a strong way. In contrast, the post-double selection estimator $\check{\alpha}$ guards against omitted variables bias which reduces the dependence of the first-order behaviour on P_n. This good behaviour under sequences P_n translates into uniform with respect to $\mathrm{P}\in\mathbf{P}$ asymptotic normality.

We should note that the post-double-selection estimator is first-order equivalent to the regression including all the controls when p is small relative to n.[15] This equivalence disappears under approximating sequences with number of controls proportional to the sample size, $p\propto n$, or greater than the sample size, $p\gg n$. It is these scenarios that motivate the use of selection as a means of regularization. In these more complicated settings the intuition from this simple $p=1$ example carries through, and the post-single selection method has a highly non-regular behaviour while the post-double selection method continues to be regular.

3. THEORY OF ESTIMATION AND INFERENCE

3.1. *Regularity conditions*

In this section, we provide 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 assumption as well as other more technical assumptions. Throughout the article, we let c, C, and q be absolute constants, and let $\ell_n\nearrow\infty,\delta_n\searrow 0$, and $\Delta_n\searrow 0$ be sequences of absolute positive constants. By absolute constants, we mean constants that are given and do not depend on the dgp $\mathrm{P}=\mathrm{P}_n$.

We assume that for each n the following condition holds on dgp $\mathrm{P}=\mathrm{P}_n$.

Condition ASTE (P): Approximate Sparse Treatment Effects. *(i) We observe $\omega_i=(y_i,d_i,z_i)$, $i=1,...,n$, where $\{\omega_i\}_{i=1}^{\infty}$ are i.n.i.d. vectors on the probability space $(\Omega,\mathcal{F},\mathrm{P})$ that obey the model (2.2)–(2.3), and the vector $x_i=P(z_i)$ is a p-dimensional dictionary of transformations of z_i, which may depend on n but not on* P. *(ii) The true parameter value α_0, which may depend*

[15] This equivalence may be a reason double-selection was previously overlooked, though there are higher-order differences between the estimator using all controls and our estimator in the case where p is small relative to n.

on P*, is bounded,* $|\alpha_0| \leqslant C$*. (iii) Functions m and g admit an approximately sparse form. Namely there exists* $s \geqslant 1$ *and* β_{m0} *and* β_{g0}*, which depend on n and* P*, such that*

$$m(z_i) = x_i'\beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{mi}^2]\}^{1/2} \leqslant C\sqrt{s/n}, \tag{3.22}$$

$$g(z_i) = x_i'\beta_{g0} + r_{gi}, \quad \|\beta_{g0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{gi}^2]\}^{1/2} \leqslant C\sqrt{s/n}. \tag{3.23}$$

(iv) The sparsity index obeys $s^2 \log^2(p \vee n)/n \leqslant \delta_n$ *and the size of the amelioration set obeys* $\widehat{s}_3 \leqslant C(1 \vee \widehat{s}_1 \vee \widehat{s}_2)$*. (v) For* $\tilde{v}_i = v_i + r_{mi}$ *and* $\tilde{\zeta}_i = \zeta_i + r_{gi}$ *we have* $|\bar{\mathrm{E}}[\tilde{v}_i^2 \tilde{\zeta}_i^2] - \bar{\mathrm{E}}[v_i^2 \zeta_i^2]| \leqslant \delta_n$*, and* $\bar{\mathrm{E}}[|\tilde{v}_i|^q + |\tilde{\zeta}_i|^q] \leqslant C$ *for some* $q > 4$*. Moreover,* $\max_{i \leqslant n} \|x_i\|_\infty^2 s n^{-1/2+2/q} \leqslant \delta_n$ *wp* $1 - \Delta_n$*.*

Comment 3.1. The approximate sparsity (iii) and rate condition (iv) are the main conditions for establishing the key inferential result. We present a number of primitive examples to show that these conditions contain standard models used in empirical research as well as more flexible models. Condition (iv) requires that the size $\widehat{s}_3$ of the amelioration set $\widehat{I}_3$ not be substantially larger than the size of the set of variables 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 dominate the number of controls selected by Lasso. This and other conditions will ensure that the total number $\widehat{s}$ of controls obeys $\widehat{s} \lesssim_P s$. We also require that $s^2 \log^2(p \vee n)/n \to 0$. Note that s is the bound on the number of regressors used by a sparse model to achieve an approximation error of order $\sqrt{s/n}$ and that the rate of convergence for the estimated coefficients would be $\sqrt{s/n}$ if we knew the identities of these s variables. Thus, the estimated function converges to the population function at a rate of $\sqrt{s/n}$ in the idealized setting where we know the identities of the relevant variables, and we would achieve an approximation rate of $o(n^{-1/4})$ under the condition that $s^2/n \to 0$ in this case. When the identities of the relevant variables are unknown, we use the stronger rate condition $s^2 \log^2(p \vee n)/n \to 0$ where the additional logarithmic term is the cost of not knowing the correct set of variables. This decrease in the rate of convergence can be substantial for large p, for example if $\log p \propto n^\gamma$ for some positive $\gamma < 1/2$. This condition can be relaxed using sample-splitting, which is done in a Supplementary Appendix. Condition (v) is simply a set of sufficient conditions for consistent estimation of the variance of the double selection estimator. If the regressors are uniformly bounded and the approximation errors are going to zero a.s., it is implied by other conditions stated below; and it can also be demonstrated under other sorts of more primitive conditions. ||

The next condition concerns the behaviour 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 behaviour of smaller submatrices. Define the minimal and maximal m-sparse eigenvalue of a semi-definite matrix M as

$$\phi_{\min}(m)[M] := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2} \text{ and } \phi_{\max}(m)[M] := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}. \tag{3.24}$$

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 (P)**: Sparse Eigenvalues.** *There is an absolute sequence* $\ell_n \to \infty$ *such that with a high probability the maximal and minimal* $\ell_n s$*-sparse eigenvalues are bounded from above and away from zero. Namely with probability at least* $1 - \Delta_n$*,*

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

where $0<\kappa'<\kappa''<\infty$ *are absolute constants.*

Comment 3.2. It is well known that Condition SE is quite plausible for many designs of interest. For instance, Condition SE holds if

(a) $(x_i)_{i=1}^n$ are i.i.d. zero-mean sub-Gaussian random vectors that have population Gram matrix $\mathrm{E}[x_ix_i']$ with minimal and maximal $s\log n$-sparse eigenvalues bounded away from zero and from above by absolute constants where $s(\log n)(\log p)/n \leqslant \delta_n \to 0$;

(b) $(x_i)_{i=1}^n$ are i.i.d. bounded zero-mean random vectors with $\|x_i\|_\infty \leqslant K_n$ a.s. such that $\mathrm{E}[x_ix_i']$ has minimal and maximal $s\log n$-sparse eigenvalues bounded from above and away from zero by absolute constants, where $K_n^2 s(\log^3 n)\{\log(p\vee n)\}/n \leqslant \delta_n \to 0$.

Claim (a) holds by Theorem 3.2 in Rudelson and Zhou (2011)[16] and claim (b) holds by Theorem 1.8 in Rudelson and Zhou (2011). Recall that a standard assumption in econometric research is to assume that the population Gram matrix $\mathrm{E}[x_ix_i']$ has eigenvalues bounded from above and away from zero, see *e.g.* Newey (1997). The conditions above allow for this and more general behaviour, requiring only that the $s\log n$ sparse eigenvalues of the population Gram matrix $\mathrm{E}[x_ix_i']$ are bounded from below and from above. ‖

The next condition imposes moment conditions on the structural errors and regressors.

Condition SM (P)**: Structural Moments.** *There are absolute constants* $0<c<C<\infty$ *and* $4<q<\infty$ *such that for* $(\tilde{y}_i,\epsilon_i)=(y_i,\zeta_i)$ *and* $(\tilde{y}_i,\epsilon_i)=(d_i,v_i)$ *the following conditions hold:*

(i) $\bar{\mathrm{E}}[|d_i|^q]\leqslant C,\;\; c\leqslant \mathrm{E}[\zeta_i^2\,|\,x_i,v_i]\leqslant C$ and $c\leqslant \mathrm{E}[v_i^2\,|\,x_i]\leqslant C$ a.s. $1\leqslant i\leqslant n$,

(ii) $\bar{\mathrm{E}}[|\epsilon_i|^q]+\bar{\mathrm{E}}[\tilde{y}_i^2]+\max_{1\leqslant j\leqslant p}\{\bar{\mathrm{E}}[x_{ij}^2\tilde{y}_i^2]+\bar{\mathrm{E}}[|x_{ij}^3\epsilon_i^3|]+1/\bar{\mathrm{E}}[x_{ij}^2]\}\leqslant C$,

(iii) $\log^3 p/n\leqslant \delta_n$,

(iv) $\max_{1\leqslant j\leqslant p}\{|(\mathbb{E}_n-\bar{\mathrm{E}})[x_{ij}^2\epsilon_i^2]|+|(\mathbb{E}_n-\bar{\mathrm{E}})[x_{ij}^2\tilde{y}_i^2]|\}+\max_{1\leqslant i\leqslant n}\|x_i\|_\infty^2\dfrac{s\log(n\vee p)}{n}\leqslant \delta_n$ wp $1-\Delta_n$.

These conditions ensure good model selection performance of feasible Lasso applied to equations (2.6) and (2.7). These conditions also allow us to invoke moderate deviation theorems for self-normalized sums from Jing *et al*. (2003) to bound some important error components.

3.2. *The main result*

The following is the main result of this article. It shows that the post-double selection estimator is root-*n* consistent and asymptotically normal. Under homoscedasticity this estimator achieves the semi-parametric efficiency bound. The result also verifies that plug-in estimates of the standard errors are consistent.

Theorem 1. (Estimation and Inference on Treatment Effects). *Let* $\{\mathrm{P}_n\}$ *be a sequence of data-generating processes. Assume conditions ASTE* (P)*, SM* (P)*, and SE* (P) *hold for* $\mathrm{P}=\mathrm{P}_n$ *for each n. Then, the post-double-Lasso estimator* $\check{\alpha}$*, constructed in the previous section, obeys as* $n\to\infty$

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0)\rightsquigarrow N(0,1),$$

where $\sigma_n^2=[\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1}$. *Moreover, the result continues to apply if* σ_n^2 *is replaced by* $\widehat{\sigma}_n^2=[\mathbb{E}_n\widehat{v}_i^2]^{-1}\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2][\mathbb{E}_n\widehat{v}_i^2]^{-1}$*, for* $\widehat{\zeta}_i:=[y_i-d_i\check{\alpha}-x_i'\check{\beta}]\{n/(n-\widehat{s}-1)\}^{1/2}$ *and*

[16] See also Zhou (2009) and Baraniuk *et al*. (2008).

$\widehat{v}_i := d_i - x_i'\widehat{\beta}$, $i=1,\ldots,n$ *where* $\widehat{\beta} \in \arg\min_{\beta}\{\mathbb{E}_n[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}$ *where* $\widehat{I}$ *is defined in (2.9).*

Comment 3.3. (Achieving the semi-parametric efficiency bound). Note that under i.i.d. sampling under P and conditional homoscedasticity, namely $\mathrm{E}[\zeta_i^2|z_i] = \mathrm{E}[\zeta_i^2]$, the asymptotic variance σ_n^2 reduces to $\mathrm{E}[v_i^2]^{-1}\mathrm{E}[\zeta_i^2]$, which is the semi-parametric efficiency bound for the partially linear model of Robinson (1988). ‖

Corollary 1. (Uniformly Valid Confidence Intervals). *Let* $\mathbf{P}_n$ *be the collection of all data-generating processes* P *for which conditions ASTE(*P*), SM (*P*), and SE (*P*) hold for given n, and let* $\mathbf{P} = \cap_{n \geqslant n_0} \mathbf{P}_n$ *be the collection of data-generating processes for which the conditions above hold for all* $n \geqslant n_0$. *Let* $c(1-\xi) = \Phi^{-1}(1-\xi/2)$. *The confidence regions based upon* $(\check{\alpha}, \widehat{\sigma}_n)$ *are valid uniformly in* $\mathrm{P} \in \mathbf{P}$*:*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}} \left|\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm c(1-\xi)\widehat{\sigma}_n/\sqrt{n}]\right) - (1-\xi)\right| = 0.$$

By exploiting both equations (2.4) and (2.5) for model selection, the post-double-selection method creates the necessary adaptivity that makes it robust to imperfect model selection. Robustness of the post-double selection method is reflected in the fact that Theorem 1 permits the data-generating process to change with n. Thus, the conclusions of the theorem are valid for a wide variety of sequences of data-generating processes which in turn define the regions $\mathbf{P}$ of uniform validity of the resulting confidence sets. In contrast, the standard post-selection method based on (2.4) produces confidence intervals that do not have close to correct coverage in many cases.

Comment 3.4. Our approach to uniformity analysis is most similar to that of Romano (2004), Theorem 4. It proceeds under triangular array asymptotics, with the sequence of dgps obeying certain constraints; then these results imply uniformity over sets of dgps that obey the constraints for all sample sizes. This approach is also similar to the classical central limit theorems for sample means under triangular arrays, and does not require the dgps to be parametrically (or otherwise tightly) specified, which then translates into uniformity of confidence regions. This approach is somewhat different in spirit to the generic uniformity analysis suggested by Andrews *et al.* (2011). ‖

Comment 3.5. (Limits of uniformity). Uniformity for inference about α_0 holds over a large class of approximately sparse models: models where both $g(\cdot)$ and $m(\cdot)$ are well approximated by $s \ll n^{1/2}$ terms so that they are both estimable at $o(n^{-1/4})$ rates. Approximate sparsity is more general than assumptions often used to justify series estimation of partially linear models, so the uniformity regions—the sets of models over which inference is valid—are substantial in that regard. We formally demonstrate this through a series of examples in Section 4.1. Of course, for every interesting class of models and any non-trivial inference method, one could find an even bigger class of models where the uniformity does not apply. For example, our approach will not work in "dense" models where $g(\cdot)$ or $m(\cdot)$ is not well approximated unless $s \gg n^{1/2}$ terms are used; such dense models would generally involve many small coefficients that decay to zero very slowly or not at all. In the series case, such a model corresponds to a deviation from smoothness towards highly non-smooth functions, for example functions generated as realized paths of a white noise process. The fact that our results do not cover such models motivates further research work on inference procedures that would provide valid inference when one considers deviations from the given class of models that are deemed important. In the simulations in

Section 4.2, we consider incorporating the ridge fit along with the other controls to be selected over using Lasso to build extra robustness against such deviations away from approximately sparse models. ||

3.3. *Inference after double selection by a generic selection method*

The conditions provided so far offer a set of sufficient conditions that are tied to the use of Lasso as the model selector. The purpose of this section is to prove that the main results apply to any other model selection method that is able to select a sparse model with good approximation properties. As in the case of Lasso, we allow for imperfect model selection. Next we state a high-level condition that summarizes a sufficient condition on the performance of a model selection method that allows the post-double selection estimator to attain good inferential properties.

Condition HLMS (P)**: High-Dimensional Linear Model Selection.** *A model selector provides possibly data-dependent sets* $\widehat{I}_1 \cup \widehat{I}_2 \subseteq \widehat{I} \subset \{1,\ldots,p\}$ *of covariate names such that, with probability* $1-\Delta_n$, $|\widehat{I}| \leqslant Cs$ *and*

$$\min_{\beta:\beta_j=0, j\notin\widehat{I}_1} \sqrt{\mathbb{E}_n[(m(z_i)-x_i'\beta)^2]} \leqslant \delta_n n^{-1/4} \text{ and } \min_{\beta:\beta_j=0, j\notin\widehat{I}_2} \sqrt{\mathbb{E}_n[(g(z_i)-x_i'\beta)^2]} \leqslant \delta_n n^{-1/4}.$$

Condition HLMS requires that with high probability the selected models are sparse and generate good approximations for the functions g and m. Examples of methods producing such models include the Dantzig selector (Candès and Tao, 2007), feasible Dantzig selector (Gautier and Tsybakov, 2011), Bridge estimator (Huang *et al.*, 2008), SCAD penalized least squares (Fan and Li, 2001), square-root-Lasso (Belloni *et al.*, 2011), and thresholded Lasso (Belloni and Chernozhukov, 2013), to name a few. We emphasize that, similarly to the previous arguments, these conditions allow for imperfect model selection. Nonetheless we note that Condition HLMS implicitly assumes that tuning parameters of the model selection procedure are set properly to achieve these conditions.

The following result establishes the inferential properties of a generic post-double-selection estimator.

Theorem 2. (Estimation and Inference on Treatment Effects under High-Level Model Selection). *Let* $\mathbf{P}_n$ *be the collection of all data-generating processes* P *for which conditions ASTE(*P*), SM (*P*), SE (*P*), and HLMS (*P*) hold for given n. (1) Then under any sequence* $\mathrm{P}_n \in \mathbf{P}_n$*, the generic post-double-selection estimator* $\check{\alpha}$ *based on* $\widehat{I}$*, as defined in (2.8), obeys*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) \rightsquigarrow N(0,1),$$

where $\sigma_n^2 = [\bar{\mathrm{E}} v_i^2]^{-1} \bar{\mathrm{E}}[v_i^2 \zeta_i^2][\bar{\mathrm{E}} v_i^2]^{-1}$. *Moreover, the result continues to apply if* σ_n^2 *is replaced by* $\widehat{\sigma}_n^2 = [\mathbb{E}_n \widehat{v}_i^2]^{-1} \mathbb{E}_n[\widehat{v}_i^2 \widehat{\zeta}_i^2][\mathbb{E}_n \widehat{v}_i^2]^{-1}$*, 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* $\widehat{\beta} \in \arg\min_\beta \{\mathbb{E}_n[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}$. *(2) Moreover, let* $\mathbf{P} = \cap_{n \geqslant n_0} \mathbf{P}_n$ *be the collection of data-generating processes for which the conditions above hold for all* $n \geqslant n_0$. *The confidence regions based upon* $(\check{\alpha}, \widehat{\sigma}_n)$ *are valid uniformly in* $\mathrm{P} \in \mathbf{P}$:

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}} \left|\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm \Phi^{-1}(1-\xi/2)\widehat{\sigma}_n/\sqrt{n}]\right) - (1-\xi)\right| = 0.$$

4. THEORETICAL AND MONTE-CARLO EXAMPLES

4.1. *Theoretical examples*

The purpose of this section is to give examples that highlight the range of the applicability of the proposed method. In these examples, we specify primitive conditions that cover certain non-parametric models and high-dimensional parametric models as corollaries. A Supplementary Appendix provides proofs for these corollaries. We emphasize that our main regularity conditions cover even more general models which combine various features of these examples. In all examples, the model is the partially linear model (2.2)–(2.3) of Section 2, however, the structure for g and m will vary across examples, and so will the assumptions on the error terms ζ_i and v_i.

4.1.1. Parametric model with fixed p. We start out with a simple example, in which the dimension p of the regressors is fixed. In practical terms this example approximates cases with p small compared to n. This simple example is important since standard post-single-selection methods fail even in this simple case. Specifically, they produce confidence intervals that are not valid uniformly in the underlying data-generating process; see Leeb and Pötscher (2008a). In contrast, the post-double-selection method produces confidence intervals that are valid uniformly in the underlying data-generating process.

Example 1. (Parametric Model with Fixed p.) Consider $(\Omega, \mathcal{A}, \mathrm{P})$ as the probability space, on which we have $\{(y_i, z_i, d_i)\}_{i=1}^{\infty}$ as i.i.d. vectors obeying the model (2.2)–(2.3) with

$$g(z_i) = \sum_{j=1}^{p} \beta_{g0j} z_{ij}, \quad m(z_i) = \sum_{j=1}^{p} \beta_{m0j} z_{ij}. \tag{4.25}$$

For estimation we use $x_i = (z_{ij}, j = 1, ..., p)'$. We assume that there are absolute constants $0 < b < B < \infty$, $q_x \geqslant q > 4$, with $4/q_x + 4/q < 1$, such that

$$b \leqslant \mathrm{E}[\zeta_i^2 \,|\, x_i, v_i], \;\; \mathrm{E}[|\zeta_i^q| \,|\, x_i, v_i] \leqslant B, \;\; b \leqslant \mathrm{E}[v_i^2 \,|\, x_i], \;\; \mathrm{E}[|v_i^q| \,|\, x_i] \leqslant B. \tag{4.26}$$

Corollary 2. (Parametric Example with Fixed p). *Let* $\mathbf{P}$ *be the collection of all regression models* P *that obey the conditions set forth in Example 1 for all n for the given constants* (p, b, B, q_x, q)*. Then, any* $\mathrm{P} \in \mathbf{P}$ *obeys Conditions ASTE* (P) *with $s = p$, SE* (P)*, and SM* (P) *for all $n \geqslant n_0$, with the constants n_0 and $(\kappa', \kappa'', c, C)$ and sequences Δ_n and δ_n in those conditions depending only on (p, b, B, q_x, q). Therefore, the conclusions of Theorem 1 hold for any sequence* $\mathrm{P}_n \in \mathbf{P}$*, and the conclusions of Corollary 1 on the uniform validity of confidence intervals apply uniformly in* $\mathrm{P} \in \mathbf{P}$.

4.1.2. Nonparametric examples. The next examples are more substantial and include infinite-dimensional models which we approximate with linear functional forms with potentially very many regressors, $p \gg n$. The key to estimation in these models is a smoothness condition that requires regression coefficients to decay at some rates. In series estimation, this condition is often directly connected to smoothness of the regression function.

Let a and A be positive constants. We shall say that a sequence of coefficients

$$\theta = \{\theta_j, j = 1, 2, ...\}$$

is a-smooth with constant A if

$$|\theta_j| \leqslant A j^{-a}, \; j = 1, 2, ...,$$

which will be denoted as $\theta \in S_A^a$. We shall say that a sequence of coefficients $\theta = \{\theta_j, j=1,2,...\}$ is a-smooth with constant A after p-rearrangement if

$$|\theta_{(j)}| \leqslant Aj^{-a},\ j=1,2,...,p,\ \ |\theta_j| \leqslant Aj^{-a},\ j=p+1,p+2,...,$$

which will be denoted as $\theta \in S_A^a(p)$, where $\{|\theta_{(j)}|, j=1,...,p\}$ denotes the decreasing rearrangement of the numbers $\{|\theta_j|, j=1,...,p\}$. Since $S_A^a \subset S_A^a(p)$, the second kind of smoothness is strictly more general than the first kind.

Here we use the term "smoothness" motivated by Fourier series analysis where smoothness of functions often translates into smoothness of the Fourier coefficients in the sense that is stated above; see, *e.g.* Kerkyacharian and Picard (1992). For example, if a function $h:[0,1]^d \mapsto \mathbb{R}$ possesses $r>0$ continuous derivatives uniformly bounded by a constant M and the terms P_j are compactly supported Daubechies wavelets, then h can be represented as $h(z) = \sum_{j=1}^{\infty} P_j(z)\theta_{hj}$, with $|\theta_{hj}| \leqslant Aj^{-r/d-1/2}$ for some constant A; see Kerkyacharian and Picard (1992). We also note that the second kind of smoothness is considerably more general than the first since it allows relatively large coefficients to appear anywhere in the series of the first p coefficients. In contrast, the first kind of smoothness only allows relatively large coefficients among the early terms in the series. Lasso-type methods are specifically designed to deal with the generalized smoothness of the second kind and perform equally well under both kinds of smoothness. In the context of series applications, smoothness of the second kind allows one to approximate functions that exhibit oscillatory phenomena or spikes, which are associated with "high-order" series terms. An example of this is the wage function example given in Belloni *et al.* (2011).

Before we proceed to other examples we discuss a way to generate sparse approximations in infinite-dimensional examples. Consider, for example, a function h that can be represented as $h(z_i) = \sum_{j=1}^{\infty} \theta_{hj} P_j(z_i)$ with coefficients $\theta_h \in S_A^a(p)$. In this case we can construct sparse approximations by simply thresholding to zero all coefficients smaller than $1/\sqrt{n}$ and with indices $j \geqslant p$. This generates a sparsity index $s \leqslant A^{\frac{1}{a}} n^{\frac{1}{2a}}$. The non-zero coefficients could be further reoptimized by using the least squares projection. More formally, given a sparsity index $s>0$, a target function $h(z_i)$, and terms $x_i = (P_j(z_i): j=1,\ldots,p)' \in \mathbb{R}^p$, we let

$$\beta_{h0} := \arg \min_{\|\beta\|_0 \leqslant s} \mathrm{E}[(h(z_i) - x_i'\beta)^2], \tag{4.27}$$

and define $x_i'\beta_{h0}$ as the best s-sparse approximation to $h(z_i)$.

Example 2. (Gaussian Model with Very Large p.) Consider $(\Omega, \mathcal{A}, \mathrm{P})$ as the probability space on which we have $\{(y_i, z_i, d_i)\}_{i=1}^{\infty}$ as i.i.d. vectors obeying the model (2.2)–(2.3) with

$$g(z_i) = \sum_{j=1}^{\infty} \theta_{gj} z_{ij}, \quad m(z_i) = \sum_{j=1}^{\infty} \theta_{mj} z_{ij}. \tag{4.28}$$

Assume that the infinite dimensional vector $w_i = (\zeta_i, v_i, z_i')'$ with j^{th} element denoted $w_i(j)$ is jointly Gaussian with covariance operator $[\mathrm{Cov}(w_i(j), w_i(k))]_{j,k \geq 1}$ that has minimal and maximal eigenvalues bounded below by an absolute constant $\underline{\kappa} > 0$ and above by an absolute constant $\overline{\kappa} < \infty$.

The main assumption that guarantees approximate sparsity is a smoothness condition on the coefficients. Let $a > 1$ and $0 < A < \infty$ be absolute constants. We require that the coefficients of the expansions in (4.28) are a-smooth with constant A after p-rearrangement, namely

$$\theta_m = (\theta_{mj}, j=1,2,...) \in S_A^a(p), \ \ \theta_g = (\theta_{gj}, j=1,2,...) \in S_A^a(p).$$

For estimation purposes we shall use $x_i=(z_{ij},j=1,...,p)'$, and assume that $|\alpha_0|\leqslant B$ and $p=p_n$ obeys

$$n^{\frac{1-a}{a}+\chi}\log^2(p\vee n)\leqslant\bar{\delta}_n,\;\; A^{1/a}n^{\frac{1}{2a}}\leqslant p\bar{\delta}_n,\;\text{ and }\;\log^3 p/n\leqslant\bar{\delta}_n,$$

for some sequence of positive constants $\bar{\delta}_n\searrow 0$ and absolute constants B and $\chi>0$.

Corollary 3. (Gaussian Nonparametric Model). *Let $\mathbf{P}_n$ be the collection of all dgp* P *that obey the conditions set forth in Example 2 for a given n and for the given constants* ($\underline{\kappa},\bar{\kappa},a,A,B,\chi$) *and sequences $p=p_n$ and $\bar{\delta}_n$. Then, as established in a Supplementary Appendix, any* $\mathrm{P}\in\mathbf{P}_n$ *obeys Conditions ASTE* (P) *with $s=A^{1/a}n^{\frac{1}{2a}}$, SE* (P)*, and SM* (P) *for all $n\geqslant n_0$, with constants n_0 and* (κ',κ'',c,C) *and sequences Δ_n and δ_n in those conditions depending only on* ($\underline{\kappa},\bar{\kappa},a,A,B,\chi$)*, p, and $\bar{\delta}_n$. Therefore, the conclusions of Theorem 1 hold for any sequence* $\mathrm{P}_n\in\mathbf{P}_n$*, and the conclusions of Corollary 1 on the uniform validity of confidence intervals apply uniformly in* $\mathrm{P}\in\mathbf{P}=\cap_{n\geqslant n_0}\mathbf{P}_n$.

Example 3. (Series Model with Very Large p.) Consider $(\Omega,\mathcal{A},\mathrm{P})$ as the probability space, on which we have $\{(y_i,z_i,d_i)\}_{i=1}^{\infty}$ as i.i.d. vectors obeying the model (2.2)–(2.3), with

$$g(z_i)=\sum_{j=1}^{\infty}\theta_{gj}P_j(z_i),\quad m(z_i)=\sum_{j=1}^{\infty}\theta_{mj}P_j(z_i),\tag{4.29}$$

where z_i has support $[0,1]^d$ with density bounded from below by constant $\underline{f}>0$ and above by constant $\bar{f}$, and $\{P_j,j=1,2,...\}$ is an orthonormal basis on $L^2[0,1]^d$ with bounded elements, *i.e.* $\max_{z\in[0,1]^d}|P_j(z)|\leqslant B$ for all $j=1,2,....$ Here all constants are taken to be absolute. Examples of such orthonormal bases include canonical trigonometric bases, *e.g.* $\{1,\sqrt{2}\cos(2\pi jz),\sqrt{2}\sin(2\pi jz):j\geqslant 1\}$ where $z\in[0,1]$.

Let $a>1$ and $0<A<\infty$ be absolute constants. We require that the coefficients of the expansions in (4.29) are a-smooth with constant A after p-rearrangement, namely

$$\theta_m=(\theta_{mj},j=1,2,...)\in S_A^a(p),\;\;\theta_g=(\theta_{gj},j=1,2,...)\in S_A^a(p).$$

For estimation purposes we shall use $x_i=(P_j(z_i),j=1,...,p)'$, and assume that $p=p_n$ obeys

$$n^{(1-a)/a}\log^2(p\vee n)\leqslant\bar{\delta}_n,\;\; A^{1/a}n^{\frac{1}{2a}}\leqslant p\bar{\delta}_n\;\text{ and }\;\log^3 p/n\leqslant\bar{\delta}_n,$$

for some sequence of absolute constants $\bar{\delta}_n\searrow 0$. We assume that there are some absolute constants $b>0$, $B<\infty$, $q>4$, with $(1-a)/a+4/q<0$, such that

$$|\alpha_0|\leqslant B,\;\; b\leqslant\mathrm{E}[\zeta_i^2\,|\,x_i,v_i],\;\;\mathrm{E}[|\zeta_i|^q\,|\,x_i,v_i]\leqslant B,\;\; b\leqslant\mathrm{E}[v_i^2\,|\,x_i],\;\;\mathrm{E}[|v_i|^q\,|\,x_i]\leqslant B.\tag{4.30}$$

Corollary 4. (Non-parametric Model with Sieve-type Regressors). *Let $\mathbf{P}_n$ be the collection of all regression models* P *that obey the conditions set forth above for a given n. Then any* $\mathrm{P}\in\mathbf{P}_n$ *obeys Conditions ASTE* (P) *with $s=A^{1/a}n^{\frac{1}{2a}}$, SE* (P)*, and SM* (P) *for all $n\geqslant n_0$, with absolute constants in those conditions depending only on* ($\underline{f},\bar{f},a,A,b,B,q$) *and $\bar{\delta}_n$. Therefore, the conclusions of Theorem 1 hold for any sequence* $\mathrm{P}_n\in\overline{\mathbf{P}_n}$*, and the conclusions of Corollary 1 on the uniform validity of confidence intervals apply uniformly in* $\mathrm{P}\in\mathbf{P}=\cap_{n\geqslant n_0}\mathbf{P}_n$.

4.2. *Monte-Carlo examples*

In this section, we examine the finite-sample properties of the post-double-selection method and compare its performance to that of a standard post-single-selection method.

All of the simulation results are based on the model

$$y_i = d_i'\alpha_0 + x_i'\theta_g + \sigma_y(d_i,x_i)\zeta_i \tag{4.31}$$

$$d_i = x_i'\theta_m + \sigma_d(x_i)v_i \tag{4.32}$$

where $(\zeta_i, v_i)' \sim N(0,I_2)$ with I_2 the 2×2 identity matrix, $p = \dim(x_i) = 200$, the covariates $x_i \sim N(0,\Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, $\alpha_0 = .5$, and the sample size n is set to 100. Inference results for all designs are based on conventional t-tests with standard errors calculated using the heteroscedasticity consistent jackknife variance estimator discussed in MacKinnon and White (1985).

We report results from three different dgp's. In the first two dgp's, we set $\theta_{g,j} = c_y\beta_{0,j}$ and $\theta_{m,j} = c_d\beta_{0,j}$ with $\beta_{0,j} = (1/j)^2$ for $j = 1,...,200$. The first dgp, which we label "Design 1", uses homoscedastic innovations with $\sigma_y(d_i,x_i) = \sigma_d(x_i) = 1$. The second dgp, "Design 2", is heteroscedastic with $\sigma_d(x_i) = \sqrt{\frac{(1+x_i'\beta_0)^2}{\mathbb{E}_n(1+x_i'\beta_0)^2}}$ and $\sigma_y(d_i,x_i) = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\mathbb{E}_n(1+\alpha_0 d_i + x_i'\beta_0)^2}}$. The constants c_y and c_d are chosen to generate desired population values for the reduced form R^2's, *i.e.* the R^2's for equations (2.6) and (2.7). For each equation, we choose c_y and c_d to generate $R^2 = 0, 0.2, 0.4, 0.6,$ and 0.8. In the heteroscedastic design, we choose c_y and c_d based on R^2 as if (4.31) and (4.32) held with v_i and ζ_i homoscedastic and label the results by R^2 as in Design 1. In the third design ("Design 3"), we use a combination of deterministic and random coefficients. For the deterministic coefficients, we set $\theta_{g,j} = c_y(1/j)^2$ for $j \le 5$ and $\theta_{m,j} = c_d(1/j)^2$ for $j \le 5$. We then generate the remaining coefficients as iid draws from $(\theta_{g,j},\theta_{m,j})' \sim N(0_{2\times 1},(1/p)I_2)$. For each equation, we choose c_y and c_d to generate $R^2 = 0, 0.2, 0.4, 0.6,$ and 0.8 in the case that all of the random coefficients are exactly equal to 0 and label the results by R^2 as in Design 1. We draw new x's, ζ's, and v's at every simulation replication, and we also generate new θ's at every simulation replication in Design 3.

We consider Designs 1 and 2 to be baseline designs. These designs do not have exact sparse representations but have coefficients that decay quickly so that approximately sparse representations are available. Design 3 is meant to introduce a modest deviation from the approximately sparse model towards a model with many small, uncorrelated coefficients. Using this we shall document that our proposed procedure still performs reasonably well, although it could be improved by incorporation of a ridge fit as one of regressors over which selection occurs.[17]

We report results for five different procedures. Two of the procedures are infeasible benchmarks: Oracle and Double-Selection Oracle estimators, which use knowledge of the true coefficient structures θ_g and θ_m and are thus unavailable in practice. The Oracle estimates α by running ordinary least squares of $y_i - x_i'\theta_g$ on d_i, and the Double-Selection Oracle estimates α by running ordinary least squares of $y_i - x_i'\theta_g$ on $d_i - x_i'\theta_m$. The other procedures we consider are

[17] In a Supplementary Appendix, we present results for 26 additional designs. The results presented in this section are sufficient to illustrate the general patterns from the larger set of results. In particular, the post-double-Lasso performed very well across all simulation designs where approximate sparsity provides a reasonable description of the dgp. Unsurprisingly, the performance deteriorates as one deviates from the smooth/approximately sparse case. However, the post-double-Lasso outperformed all other feasible procedures considered in all designs.

feasible. One procedure is the standard post-single selection estimator—the Post-Lasso—which applies Lasso to equation (4.31) without penalizing α, the coefficient on d_i, to select additional control variables from among x. Estimates of α_0 are then obtained by OLS regression of y_i on d_i and the set of additional controls selected in the Lasso step and inference using the Post-Lasso estimator proceeds using conventional heteroscedasticity robust OLS inference from this regression. Post-Double-Selection or Post-Double-Lasso is the feasible procedure advocated in this paper. We run Lasso of y_i on x_i to select a set of predictors for y_i and run Lasso of d_i on x_i to select a set of predictors for d_i. α_0 is then estimated by running OLS regression of y_i on d_i and the union of the sets of regressors selected in the two Lasso runs, and inference is simply the usual heteroscedasticity robust OLS inference from this regression.[18] Post-Double-Selection + Ridge is an *ad hoc* variant of Post-Double-Selection in which we add the ridge fit from equation (4.32) as an additional potential regressor that may be selected by Lasso. The ridge fit for d_i is $x_i'(X'X+\lambda_d I_p)^{-1}X'D$ where λ_d is obtained by 10-fold cross-validation. This procedure is motivated by a desire to add further robustness in the case that many small coefficients are suspected and zeroing out these small coefficients may be undesirable. Further exploration of procedures that perform well, both theoretically and in simulations, in the presence of many small coefficients is an interesting avenue for additional research.

We start by summarizing results in Table 1 for $(R_y^2, R_d^2) = (0,0.2),(0,0.8),(0.8,0.2)$, and $(0.8,0.8)$ where R_y^2 is the population R^2 from regressing y on x (Structure R^2) and R_d^2 is the population R^2 from regressing d on x (First Stage R^2). We report root-mean-square-error (RMSE) for estimating α_0 and size of 5% level tests (Rej. Rate). As should be the case, the Oracle and Double-Selection Oracle, which are reported to provide the performance of an infeasible benchmark, perform well relative to the feasible procedures across the three designs. We do see that the feasible Post-Double-Selection procedures perform similarly to the Double-Selection Oracle without relying on *ex ante* knowledge of the coefficients that go in to the control functions, θ_g and θ_m. On the other hand, the Post-Lasso procedure generally does not perform as well as Post-Double-Selection and is very sensitive to the value of R_d^2. While Post-Lasso performs adequately when R_d^2 is small, its performance deteriorates quickly as R_d^2 increases. This lack of robustness of traditional variable selection methods such as Lasso which were designed with forecasting, not inference about treatment effects, in mind is the chief motivation for our advocating the Post-Double-Selection procedure when trying to infer structural or treatment parameters.

We provide further details about the performance of the feasible estimators in Figures 2, 3, and 4 which plot size of 5% level tests, bias, and standard deviation for the Post-Lasso, Double-Selection (DS), and Double-Selection Oracle (DS Oracle) estimators of the treatment effect across the full set of R^2 values considered. Figure 2, 3, and 4 respectively report the results from Design 1, 2, and 3. The figures are plotted with the same scale to aid comparability, and rejection frequencies for Post-Lasso were censored at 0.5 for readability. Perhaps the most striking feature of the figures is the poor performance of the Post-Lasso estimator. The Post-Lasso estimator performs poorly in terms of size of tests across many different R^2 combinations and can have an order of magnitude more bias than the corresponding Post-Double-Selection estimator. The behavior of Post-Lasso is quite non-uniform across R^2 combinations, and Post-Lasso does not reliably control size distortions or bias except in the case where the controls are

[18] All Lasso estimates require the choice of penalty parameter and loadings. We use the iterative procedure of Belloni *et al*. (2012) to estimate the penalty loadings using a maximum of five iterations. We set the penalty parameter according to equation (18) in Belloni and Chernozhukov (2011b) with $c=1.1$, $\alpha=.05$ and $\sigma=1$ since the variance of the score is accounted for in the penalty loadings.

TABLE 1
Simulation Results for Selected R^2 Values

	First Stage $R^2=0.2$ Structure $R^2=0$		First Stage $R^2=0.2$ Structure $R^2=0.8$		First Stage $R^2=0.8$ Structure $R^2=0$		First Stage $R^2=0.8$ Structure $R^2=0.8$	
Estimation procedure	RMSE	Rej. Rate	RMSE	Rej. Rate	RMSE	Rej. Rate	RMSE	Rej. Rate
A. Design 1. Quadratic decay								
Oracle	0.090	0.048	0.090	0.048	0.045	0.057	0.045	0.057
Double-selection oracle	0.102	0.050	0.102	0.050	0.143	0.047	0.143	0.047
Post-Lasso	0.137	0.205	0.110	0.064	0.402	0.987	0.489	0.974
Double-selection	0.107	0.063	0.107	0.058	0.109	0.074	0.104	0.062
Double-selection + ridge	0.260	0.064	0.256	0.055	0.132	0.049	0.130	0.050
B. Design 2. Quadratic decay with heteroscedasticity								
Oracle	0.139	0.060	0.139	0.060	0.066	0.062	0.066	0.062
Double-selection Oracle	0.169	0.072	0.169	0.072	0.225	0.085	0.225	0.085
Post-Lasso	0.175	0.139	0.178	0.097	0.409	0.994	0.501	0.993
Double-selection	0.165	0.098	0.167	0.081	0.162	0.082	0.165	0.083
Double-selection + ridge	0.308	0.060	0.290	0.058	0.183	0.064	0.185	0.075
C. Design 3. Quadratic decay with random coefficients								
Oracle	0.070	0.055	0.070	0.055	0.041	0.060	0.041	0.060
Double-selection oracle	0.114	0.056	0.114	0.056	0.151	0.058	0.151	0.058
Post-Lasso	0.105	0.082	0.131	0.133	0.329	0.940	0.435	0.953
Double-selection	0.109	0.055	0.118	0.075	0.105	0.056	0.117	0.086
Double-selection + ridge	0.227	0.040	0.230	0.035	0.151	0.054	0.153	0.057

Note: The table reports root-mean-square-error (RMSE) rejection rates for 5% level tests (Rej. Rate) from a Monte Carlo simulation experiment. Results are based on 1000 simulation replications. Data in Panels A and B are based on models with coefficients that decay quadratically, and the data in Panel C are based on a with five quadratically decaying coefficients and 95 random coefficients. Further details about the simulation models are provided in the text as are details about the estimation procedures. Rejection rates are for t-tests of the null hypothesis that the structural coefficient is equal to the true population value and are formed using jack-knife standard errors that are robust to heteroscedasticity; see MacKinnon and White (1985).

uncorrelated with the treatment (where First-Stage R^2 equals 0) and thus ignorable. In contrast, the Post-Double-Selection estimator performs relatively well across the full range of R^2 combinations considered. The Post-Double-Selection estimator's performance is also quite similar to that of the infeasible Double-Selection Oracle across the majority of R^2 values considered. Comparing across Figures 2 and 3, we see that size distortions for both the Post-Double-Selection estimator and the Double-Selection Oracle are somewhat larger in the presence of heteroscedasticity but that the basic patterns are more-or-less the same across the two figures. Looking at Figure 4, we also see that the addition of small independent random coefficients results in somewhat larger size distortions for the Post-Double-Selection estimator than in the other homoscedastic design, Design 1, though the procedure still performs relatively well.

In the final figure, Figure 5, we compare the performance of the Post-Double-Selection procedure to the *ad hoc* Post-Double-Selection procedure that selects among the original set of variables augmented with the ridge fit obtained from equation (4.32). We see that the addition of this variable does add robustness relative to Post-Double-Selection using only the raw controls in the sense of producing tests that tend to have size closer to the nominal level. This additional robustness is a good feature, though it comes at the cost of increased RMSE which is especially prominent for small values of the first-stage R^2.

The simulation results are favourable to the Post-Double-Selection estimator. In the simulations, we see that the Post-Double-Selection procedure provides an estimator of a treatment effect in the presence of a large number of potential confounding variables that performs similarly to the infeasible estimator that knows the values of the coefficients on all of the confounding

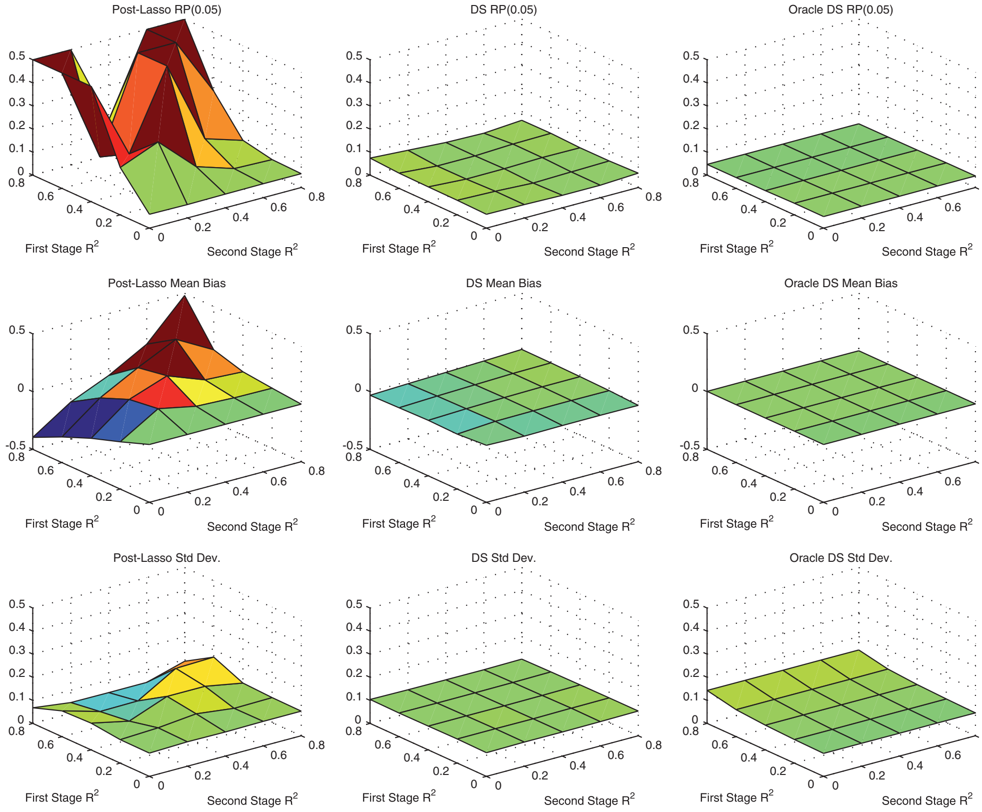

FIGURE 2

This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 1 of the simulation study that has quadratically decaying coefficients and homoscedasticity. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and an infeasible OLS estimator that relies on knowledge of the true values of the coefficients in equations (2.6) and (2.7). Reduced form and first stage R^2 correspond to the population R^2 of (2.6) and (2.7), respectively. Note that rejection frequencies are censored at 0.5.

variables. Overall, the simulation evidence supports our theoretical results and suggests that the proposed Post-Double-Selection procedure can be a useful tool to researchers doing structural estimation in the presence of many potential confounding variables. It also shows, as a contrast, that the standard Post-Single-Selection procedure provides poor inference and therefore is not a reliable tool to these researchers.

5. GENERALIZATIONS AND HETEROGENEOUS TREATMENT EFFECTS

In order to discuss generalizations in a simple manner, we assume i.i.d sampling and no approximation errors for a moment (*i.e.* we let $g(z_i)=x_i'\beta_{g0}$ and $m(z_i)=x_i'\beta_{m0}$, where $x_i=P(z_i)$). In the development of results for the partially linear model, we implicitly considered a moment condition for the target parameter α_0 given by

$$\mathrm{E}[\psi(y_i-d_i\alpha_0-x_i'\beta)v_i]=0, \tag{5.33}$$

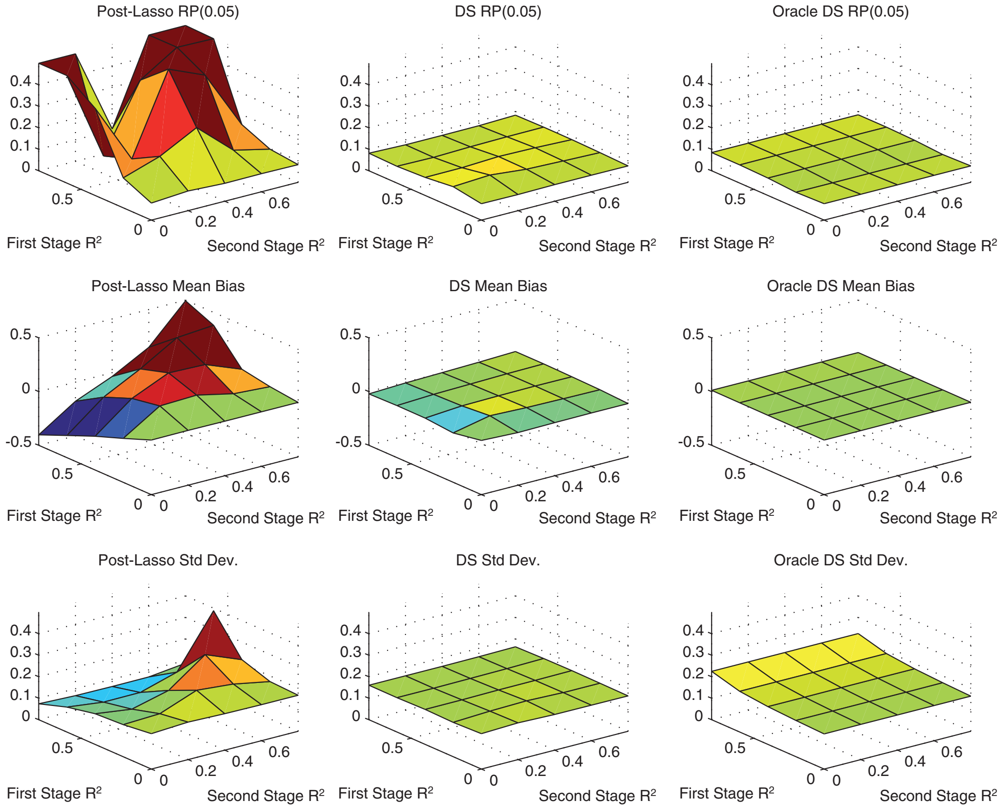

FIGURE 3

This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 2 of the simulation study which has quadratically decaying coefficients and heteroscedasticity. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and an infeasible OLS estimator that relies on knowledge of the true values of the coefficients in equations (2.6) and (2.7). Reduced form and first stage R^2 correspond to the population R^2 of (2.6) and (2.7), respectively. Note that rejection frequencies are censored at 0.5.

where $x_i = P(z_i)$, $\psi(u) = u$, and v_i are measurable functions of z_i, the "instruments". We selected the instruments v_i such that the equation is first-order insensitive to the parameter β at $\beta = \beta_{g0}$:

$$\frac{\partial}{\partial\beta}\mathrm{E}[\psi(y_i - d_i\alpha_0 - x_i'\beta)v_i]\Big|_{\beta=\beta_{g0}} = 0. \tag{5.34}$$

Note that when $\psi(u) = u$, the "instrument" $v_i = d_i - m(z_i)$ precisely implements this condition. If (5.34) holds, the estimator of α_0 based upon the sample analogue of (5.33) gets "immunized" against non-regular estimation of β_0, for example, via a post-selection procedure or other regularized estimators. Such immunization ideas are in fact behind the classical Frisch-Waugh and Robinson (1988) partialling out technique in the linear setting and the Neyman (1979)'s $C(\alpha)$ test in the nonlinear setting. One way to view our contribution is as a recognition of the importance of this immunization in the context of post-selection inference leading to thinking

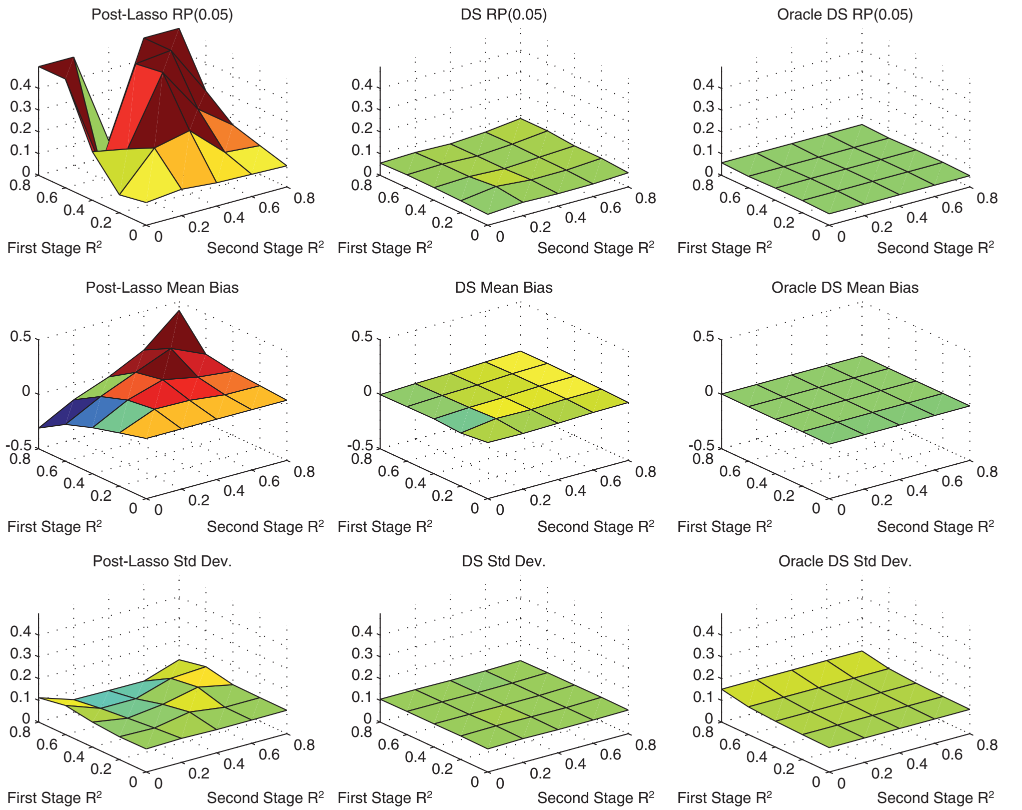

FIGURE 4

This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 3 of the simulation study that has five quadratically decaying coefficients and 95 Gaussian random coefficients. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and an infeasible OLS estimator that relies on knowledge of the true values of the coefficients in equations (2.6) and (2.7). Reduced form and first stage R^2 correspond to what would be the population R^2 of (2.6) and (2.7) if all of the random coefficients were equal to zero. Note that rejection frequencies are censored at 0.5.

about a post-selection approach to inference on the target parameter that uses this condition.[19] Generalizations to non-linear models, where ψ is non-linear and can correspond to a likelihood score or quantile check function are given in Belloni *et al*. (2013a) and Belloni *et al*. (2013b); in these generalizations, achieving (5.34) is also critical.

In the context of the present paper, an important generalization is the estimation of average treatment effects (ATE) when treatment effects are fully heterogeneous and the treatment variable is binary, $d_i \in \{0,1\}$. We consider i.i.d. vectors $\{(y_i, d_i, z_i)\}_{i=1}^{\infty}$ on the probability space $(\Omega, \mathcal{A}, \mathrm{P})$,

[19] To the best of our knowledge, Belloni *et al*. (2010) and Belloni *et al*. (2012) were the first to use this immunization/orthoganility property in the $p \gg n$ setup, in the context of performing inference on low-dimensional parameters in the instrumental regression, where the nuisance function being estimated via regularization or post-selection is the optimal instrument. There the orthogonality property was used to establish the asymptotic normality and $\sqrt{n}$ consistency of the resulting estimator under rich sequence of data-generating processes P_n, which translates to uniformity over suitably defined regions.

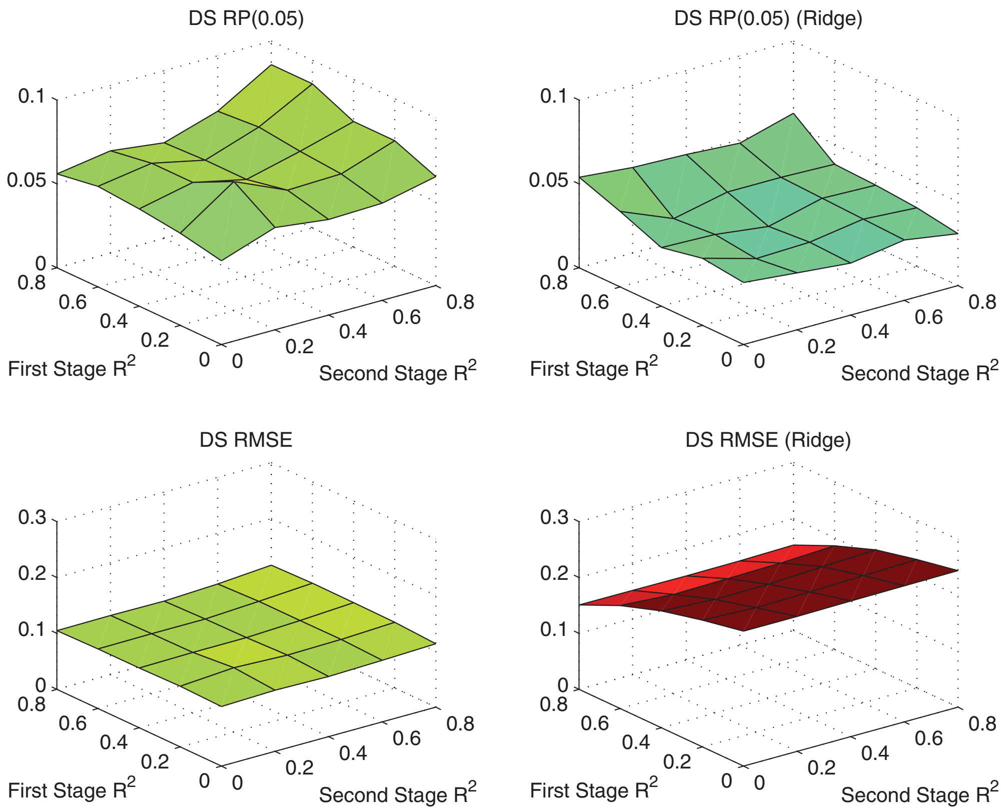

FIGURE 5

This figure presents rejection frequencies for 5% level tests and RMSE's for estimating the treatment effect from Design 3 of the simulation study that has five quadratically decaying coefficients and 95 Gaussian random coefficients. Results in the first column are for the proposed double selection procedure, and the results in the second column are for the proposed double selection procedure when the ridge fit from (2.6) is added as an additional potential control. Reduced form and first stage R^2 correspond to what would be the population R^2 of (2.6) and (2.7) if all of the random coefficients were equal to zero. Note that the vertical axis on the rejection frequency graph is from 0 to 0.1.

and suppose the the outcome and propensity equations are

$$y_i = g(d_i, z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0, \tag{5.35}$$

$$d_i = m(z_i) + v_i, \quad \mathrm{E}[v_i \mid z_i] = 0. \tag{5.36}$$

A common target parameter of interest in this model is the average treatment effect (ATE),

$$\mathrm{E}[g(1, z_i) - g(0, z_i)].$$

Another common target parameter is the average treatment effect for the treated (ATT) $\mathrm{E}[g(1, z_i) - g(0, z_i) | d_i = 1]$. In this model, d_i is not additively separable, and we no longer have a partially linear model, but our analysis easily extends to this case.

The confounding factors z_i affect the policy variable via the propensity score $m(z_i)$ and the outcome variable via the function $g(d_i, z_i)$. Both of these functions are unknown and potentially complicated. We use control terms $x_i = P(z_i)$ in approximating $g(d_i, z_i)$ and $m(z_i)$. Specifically, we write (2.2) and (2.3) as

$$y_i = \underbrace{\tilde{x}_i'\beta_{g0} + r_{gi}}_{g(d_i,z_i)} + \zeta_i, \quad d_i = \underbrace{\Lambda(x_i'\beta_{m0}) + r_{mi}}_{m(z_i)} + v_i, \tag{5.37}$$

where r_{gi} and r_{mi} are approximation errors;

$$\tilde{x}_i := (d_i x_i', (1-d_i)x_i')'; \quad \beta_{g0} := (\beta_{g0,1}', \beta_{g0,0}')'; \tag{5.38}$$

$x_i'\beta_{g0,1}$, $x_i'\beta_{g0,0}$, and $\Lambda(x_i'\beta_{m0})$ are approximations to $g(1,z_i)$, $g(0,z_i)$, and $m(z_i)$; and $\Lambda(u) = u$ for the case of linear link and $\Lambda(u) = e^u/(1+e^u)$ for the case of the logistic link. In order to allow for a flexible specification and incorporation of pertinent confounding factors, the dimension of the vector of controls, $x_i = P(z_i)$, can be large relative to the sample size. We use post-Lasso estimators $\widehat{g}(d,z) = \tilde{x}'\widehat{\beta}_{g0}$ and $\widehat{m}(z) = \Lambda(x'\widehat{\beta}_{m0})$ of functions $g(d,z)$ and $m(z)$ based upon equations (5.37)-(5.37). In case of using the logistic link Λ, Lasso for logistic regression is as defined in van de Geer (2008) and Bach (2010), and the associated post-Lasso estimators are as defined in Belloni *et al*. (2013b).

Identification of the true value α_0 of the target parameter α, either the ATE or ATT, will be based on a moment condition of type

$$\mathrm{E}[\varphi(\alpha, \omega_i, h_0(z_i))] = 0, \tag{5.39}$$

and the "post-double-selection estimator" $\check{\alpha}$ will be based on the finite-sample analogue of the moment condition

$$\mathbb{E}_n\big[\varphi(\check{\alpha}, \omega_i, \widehat{h}_0(\omega_i))\big] = 0, \tag{5.40}$$

where $\omega_i = (y_i, d_i, z_i)'$ where φ, h_0, $\widehat{h}_0$ differ depending on the target and are defined below.

For estimation of the ATE, we employ

$$\begin{gathered}\varphi(\alpha, \omega_i, h(z_i)) := \alpha - \frac{d_i(y_i - h_2(z_i))}{h_3(z_i)} - \frac{(1-d_i)(y_i - h_1(z_i)))}{1 - h_3(z_i)} - h_1(z_i) - h_2(z_i), \\ h_0(z_i) := (g(0,z_i), g(1,z_i), m(z_i))', \ \ \widehat{h}_0(z_i) := (\widehat{g}(0,z_i), \widehat{g}(1,z_i), \widehat{m}(z_i))',\end{gathered} \tag{5.41}$$

where $h(z_i) := (h_j(z_i))_{j=1}^3$ is the nuisance parameter, consisting of measurable functions mapping the support of z_i to $\mathbb{R} \times \mathbb{R} \times (0,1)$. The true value of this parameter is given above by $h_0(z_i)$, and the estimators $\widehat{g}(0,z_i), \widehat{g}(1,z_i), \widehat{m}(z_i)$ are post-Lasso estimators of functions g and m based upon equations (5.37)-(5.37). The function $\varphi(\alpha_0, \omega_i, h_0(z_i))$ is the efficient influence function of Hahn (1998) for estimating ATE. Similarly, we use

$$\begin{gathered}\varphi(\alpha, \omega_i, h(z_i)) = \frac{d_i(y_i - h_2(z_i))}{h_4} - \frac{h_3(z_i)(1-d_i)(y_i - h_1(z_i))}{(1 - h_3(z_i))h_4} + \frac{d_i(h_2(z_i) - h_1(z_i))}{h_4} - \alpha\frac{d_i}{h_4}, \\ h_0(z_i) = (g(0,z_i), g(1,z_i), m(z_i), \mathrm{E}[d_i])', \ \ \widehat{h}_0(z_i) = (\widehat{g}(0,z_i), \widehat{g}(1,z_i), \widehat{m}(z_i), \mathbb{E}_n[d_i]),\end{gathered} \tag{5.42}$$

for estimation of the ATT. Again, $\varphi(\alpha_0, \omega_i, h_0(z_i))$ is the efficient influence function of Hahn (1998) for estimating the ATT.

It is straightforward to check that in either of the previous cases, the following "immunization" property holds:[20]

$$\frac{\partial}{\partial h}\mathrm{E}[\varphi(\alpha_0,\omega_i,h(z_i))]_{h=h_0}=0, \tag{5.43}$$

where the left side denotes the pathwise derivative operator with respect to the functional parameter h at $h=h_0$. This is a generalization of the condition (5.34). Thus, we reduce dependence on the estimated values of $h_0(z_i)$ by using Hahn (1998)'s efficient influence functions just as in the partially linear case. As before, this property suggests that one can use the selection approach to regularization in order to estimate the parameter of interest α_0.

In what follows, we use $\|w_i\|_{\mathrm{P},q}$ to denote the $L^q(\mathrm{P})$ norm of a random variable w_i with law determined by P, and $\|w_i\|_{\mathbb{P}_n,q}$ to denote the empirical $L^q(\mathbb{P}_n)$ norm of a random variable with law determined by the empirical measure $\mathbb{P}_n=n^{-1}\sum_{i=1}^n\delta_{w_i}$, *i.e.* $\|w_i\|_{\mathbb{P}_n,q}=(n^{-1}\sum_{i=1}^n\|w_i\|^q)^{1/q}$. Consider fixed sequences of positive numbers $\delta_n\searrow 0$ and $\Delta_n\searrow 0$ and constants $C>0, c>0, 1/2>c'>0$ which will not vary with P.

Condition HTE (P). **Heterogeneous Treatment Effects.** *Consider i.i.d. vectors* $\{(y_i,d_i,z_i)\}_{i=1}^{\infty}$ *on the probability space* $(\Omega,\mathcal{A},\mathrm{P})$*, such that equations (5.35)-(5.37) holds, with* $d_i\in\{0,1\}$*. (i) Approximation errors satisfy* $\|r_{gi}\|_{\mathrm{P},2}\leqslant\delta_n n^{-1/4}$*,* $\|r_{gi}\|_{\mathrm{P},\infty}\leqslant\delta_n$*, and* $\|r_{mi}\|_{\mathrm{P},2}\leqslant\delta_n n^{-1/4}$*,* $\|r_{mi}\|_{\mathrm{P},\infty}\leqslant\delta_n$*. (ii) With* P*-probability no less than* $1-\Delta_n$ *and* K_n *defined below, estimation errors satisfy* $\|\tilde{x}_i'(\widehat{\beta}_g-\beta_{g0})\|_{\mathbb{P}_n,2}\leqslant\delta_n n^{-1/4}$*,* $\|x_i'(\widehat{\beta}_m-\beta_{m0})\|_{\mathbb{P}_n,2}\leqslant\delta_n n^{-1/4}$*,* $K_n\|\widehat{\beta}_m-\beta_m\|_1\leqslant\delta_n$*,* $K_n\|\widehat{\beta}_m-\beta_{m0}\|_1\leqslant\delta_n$*, estimators and approximations are sparse, namely* $\|\widehat{\beta}_g\|_0\leqslant Cs$*,* $\|\widehat{\beta}_m\|_0\leqslant Cs$*,* $\|\beta_{g0}\|_0\leqslant Cs$*, and* $\|\beta_{m0}\|_0\leqslant Cs$*, and the empirical and populations norms are equivalent on sparse subsets, namely* $\sup_{\|\delta\|_0\leqslant 2Cs}\left|\|\tilde{x}_i'\delta\|_{\mathbb{P}_n,2}/\|\tilde{x}_i'\delta\|_{\mathrm{P},2}-1\right|\leqslant\delta_n$*. (iii) The following boundedness conditions hold:* $\|x_{ij}||_{\mathrm{P},\infty}\leqslant K_n$ *for each j,* $\|g\|_{\mathrm{P},\infty}\leqslant C$*,* $\|y_i\|_{\mathrm{P},\infty}\leqslant C$*,* $\mathrm{P}(c'\leqslant m(z_i)\leqslant 1-c')=1$*, and* $\|\zeta_i^2\|_{\mathrm{P},2}\geqslant c$*. (iv) The sparsity index obeys the following growth condition,* $(s\log(p\vee n))^2/n\leqslant\delta_n$*.*

These conditions are simple high-level conditions that encode both the approximate sparsity of the models as well as impose some reasonable behaviour on the sparse estimators of m and g. These conditions are implied by other more primitive conditions in the literature; see van de Geer (2008) and Belloni *et al.* (2012). Sufficient conditions for the equivalence between population and empirical sparse eigenvalues are given in Lemmas SA.7 and SA.8 in the Supplementary Appendix. The boundedness conditions are made to simplify arguments, and they could be removed at the cost of more complicated proofs and more stringent side conditions.

Theorem 3. (Uniform Post-Double Selection Inference on ATE and ATT). *(1) Suppose that the ATE* $\alpha_0=\mathrm{E}[g(1,z_i)-g(0,z_i)]$ *is the target and we use the estimator* $\check{\alpha}$ *and other notations defined via (5.40) and (5.41). (2) Or, alternatively, suppose that the ATT* $\alpha_0=\mathrm{E}[g(1,z_i)-g(0,z_i)|d_i=1]$ *is the target and we use the estimator* $\check{\alpha}$ *and other notations defined via (5.40) and (5.42). Consider the set* $\mathbf{P}_n$ *of data generating processes* P *such that equations (5.35)–(5.36) and Condition HTE (P) holds for given n. Then in either case, under any sequence* $\mathrm{P}\in\mathbf{P}_n$*,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0)\rightsquigarrow N(0,1),\ \ \sigma_n^2=\mathrm{E}[\varphi^2(\alpha_0,\omega_i,h_0(z_i))]. \tag{5.44}$$

The result continues to hold with σ_n^2 *replaced by* $\widehat{\sigma}_n^2:=\mathbb{E}_n[\varphi^2(\check{\alpha},\omega_i,\widehat{h}_0(z_i))]$*. Moreover, the confidence regions based upon post-double selection estimator* $\check{\alpha}$ *have uniform asymptotic*

[20] In fact, some higher order derivatives also vanish. This higher order property can be exploited in conjunction with sample-splitting to relax requirements on s. We do not discuss the details of such an approach here for brevity.

validity: $\lim_{n\to\infty}\sup_{\mathrm{P}\in\mathbf{P}}|\mathrm{P}\big(\alpha_0\in[\check{\alpha}\pm\Phi^{-1}(1-\xi/2)\widehat{\sigma}_n/\sqrt{n}]\big)-(1-\xi)|=0$*, where* $\mathbf{P}=\cap_{n\geqslant n_0}\mathbf{P}_n$ *is the collection of data-generating processes* P *for which HTE(*P*) holds for all* $n\geqslant n_0$.

Comment 5.1. These results contribute to recent results on estimation of the ATE and its variants[21] in Cattaneo (2010), who considers this problem in a series framework where the number of series terms obeys $p^2/n\to 0$, and in Rothe and Firpo (2013), who provide results based on kernel estimators. These approaches are very useful but do not target "data-rich-environments" and do not study uniformity. Our framework allows for consideration of a large number of series terms, potentially much larger than the sample size, but requires that a relatively small number of these terms are needed through the sparsity condition $s^2K_n(\log(p\vee n))^2/n\to 0$. Our framework also covers semi-parametric models with a large number of raw regressors x_{ij} as long as $|x_{ij}|\leqslant K_n$ where K_n does not grow too quickly. Finally, we establish validity of our inferential results uniformly in P. We also refer the reader to the independent work by Farrell (2013), who develops group-Lasso methods for estimating ATE in the heterogeneous effects framework. ||

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 estimate 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 Donohue III and Levitt'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 article.

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 15–25 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 favourable 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 clear 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 overall state level abortion and crime rates. It is also important to control flexibly for aggregate trends. For example, it could be the case that national crime rates were falling over some 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

[21] Further results can be found in Belloni *et al*. (2013).

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 for state-level crime rates running from 1985 to 1997 in which they condition on a number of these factors. Their basic specification is

$$y_{cit}=\alpha_c a_{cit}+w_{it}'\beta_c+\delta_{ci}+\gamma_{ct}+\varepsilon_{cit} \tag{6.45}$$

where i indexes states, t indexes times, $c\in\{\text{violent, property, murder}\}$ indexes type of crime, δ_{ci} are state-specific effects that control for any time-invariant state-specific characteristics, γ_{ct} 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,[22] 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 unemployment rate, per-capita income, 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 (6.45) 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 (6.45). 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.[23] Thus, we will estimate models of the form

$$y_{cit}-y_{cit-1}=\alpha_c(a_{cit}-a_{cit-1})+z_{cit}'\kappa_c+g_{ct}+\eta_{cit}. \tag{6.46}$$

where g_{ct} are time effects. 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 (6.45) 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 (6.45) are given in the first and second row of Table 2, respectively.

[22] 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 0.6 times the abortion rate in state i at time $t-18$ plus 0.4 times the abortion rate in state i at time $t-19$. See Donohue III and Levitt (2001) for further detail and exact construction methods.

[23] Part of the motivation for considering first-differences is that our theoretical results are for independent data. For both violent crime and property crime, this assumption seems like a better approximation in differences than in levels. The first three estimated autocorrelations of the first-difference residuals from the baseline specification using only the controls from Donohue III and Levitt (2001) based on violent crime, property crime, and murder are respectively (0.0155, 0.0574, −0.0487), (−0.0736, 0.0651, 0.0540), and (−0.3954, −0.0813, 0.0066). Discussion of results obtained estimating the model in levels and using fixed effects are available in a Supplementary Appendix. Extending the formal results to accommodate dependence would be a useful extension for future work.

TABLE 2
Estimated Effects of Abortion on Crime Rates

	Violent crime		Property crime		Murder	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
A. Donohue III and Levitt (2001) Table IV						
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	0.014	0.719	−0.195	0.225	2.343	2.798
Post-double-selection	−0.104	0.107	−0.030	0.055	−0.125	0.151
Post-double-selection+	−0.082	0.106	−0.031	0.057	−0.068	0.200

Note: The table displays the estimated coefficient on the abortion rate, “Effect”, and its estimated standard error. Numbers in the first row are taken from Donohue III 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 labelled “First-Difference” are obtained using the same controls as in the first row. Estimates in the row labelled “All Controls” use 284 control variables as discussed in the text. Estimates in the row “Post-Double-Selection” use the variable selection technique developed in this article to search among the set of 284 potential controls. Estimates in the row “Post-Double-Selection+” use the variables selected by the procedure of this article augmented with the set of variables from Donohue III and Levitt (2001).

Our main point of departure from Donohue III and Levitt (2001) is that we allow for a much richer set z_{cit} than allowed for in w_{it} in model (6.45). Our z_{cit} includes higher order terms and interactions of the control variables defined above. In addition, we put initial conditions and initial differences of w_{it} and a_{cit} and within-state averages of w_{it} into our vector of controls z_{cit}. This addition allows for the 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_{cit} as well as interactions of this quadratic trend with control variables. This gives us a set of 284 control variables to select among in addition to the 12 time effects that we include in every model.[24]

Note that interpreting estimates of the effect of abortion from model (6.45) 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 284 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 284 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 article offer one resolution to this tension. The assumed sparse structure maintains that there is a small enough set

[24] The exact identities of the 284 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_{cit}, the within-state averages of w_{it}, and interactions of these variables with a quadratic trend.

of variables that one could potentially learn about the treatment, but does add 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.

In the abortion example, we use the post-double-selection estimator defined in Section 2.2 for each of our dependent variables.[25] For violent crime, eight variables are selected in the abortion equation,[26] and no variables are selected in the crime equation. For property crime, nine variables are selected in the abortion equation,[27] and three are selected in the crime equation.[28] For murder, nine variables are selected in the abortion equation,[29] 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 284 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 eight, twelve, and nine controls selected by the double-post-Lasso procedure for violent crime, property crime, and murder respectively. All of these estimates for the effect of abortion on crime rates are quite imprecise, producing 95% confidence intervals that encompass large positive and negative values. Note that the post-double-Lasso produces models that are not of vastly different size than the "intuitive" model (6.45). As a final check, we also report results that include all of the original variables from (6.45) in the amelioration set in the fifth row of the table. These results show that the conclusions made from using only the variable selection procedure do not qualitatively change when the variables used in the original Donohue III and Levitt (2001) are added to the equation. For a quick benchmark relative to the simulation examples, we note that the R^2 obtained by regressing the crime rate on the selected variables are 0.0251, 0.1179, and 0.0039 for violent crime, property crime, and the murder rate respectively and that the R^2's from regressing the abortion rate on the selected variables are 0.8420, 0.6116, and 0.7781 for violent crime, property crime, and the murder rate respectively. These values correspond to regions of the R^2 space considered in the simulation where the double selection procedure substantially outperformed simple post-single selection procedures.

It is 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

[25] Implementation requires selection of a penalty parameter and loadings. We estimate the loadings using the iterative procedure proposed in Belloni *et al*. (2012) with 100 as the maximum number of iterations. For each model, the iterative procedure converges after 21 or fewer iterations. We set the penalty parameter according to (2.12) with $c = 1.1$ and $\gamma = 0.05$.

[26] The selected variables are lagged prisoners per capita, the lagged unemployment rate, the initial change in beer consumption interacted with a linear trend, the initial change in income squared interacted with a linear trend, the within-state mean of income, the within-state mean of lagged prisoners per capita interacted with a linear trend, the within-state mean of income interacted with a linear trend, and the initial level of the abortion rate.

[27] The selected variables are lagged prisoners per capita, lagged income, the initial change in income, the initial level of income, the initial change in beer consumption interacted with a linear trend, the initial change of income squared interacted with a linear trend, the within-state average of income, the within-state average of income interacted with a linear trend, and the initial level of abortion.

[28] The three variables are the initial level income squared interacted with a linear trend, the within-state average of AFDC generosity, and the within-state average of AFDC generosity squared.

[29] The selected variables are lagged prisoners per capita, lagged unemployment, the initial change in unemployment squared, the initial level of prisoners per capita interacted with a linear trend, the initial change in beer consumption interacted with t^2, the within-state average of the number of prisoners per capita interacted with a linear trend, the within-state average of income interacted with a linear trend, the initial level of the abortion rate, and the initial level of the abortion rate interacted with a linear trend.

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. We also see that we cannot precisely determine the effect of the abortion rate on crime rates once one accounts for these 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. Interestingly, a similar conclusion is given in Foote and Goetz (2008) based on an intuitive argument.

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 for the differences.

APPENDIX A. ITERATED LASSO ESTIMATION

Feasible implementation of Lasso under heteroscedasticity requires a choice of penalty parameter λ and estimation of penalty loadings (2.12). λ depends only on the known p and n and the researcher specified c and γ. In all examples, we use $c = 1.1$ and $\gamma = 0.05$. In this appendix, we state algorithms for estimating the penalty loadings.

Let I_0 be an initial set of regressors with a bounded number of elements, including for example 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 penalty loadings using Post-Lasso is as follows:

Algorithm 1. (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 $l_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widetilde{\beta})^2]}\sqrt{n/(n-\widehat{s})}$. (3) If $\max_{1 \leqslant j \leqslant p} |\widehat{l}_{j,k} - \widehat{l}_{j,k+1}| \leqslant \nu$ or $k > K$, set the loadings to $\widehat{l}_{j,k+1}$, $j = 1, \ldots, p$ and stop; otherwise, set $k \leftarrow k+1$ and go to (1).*

APPENDIX B. AUXILIARY RESULTS ON MODEL SELECTION VIA LASSO AND POST-LASSO

The post-double-selection estimator applies the least squares estimator to the union of variables selected for equations (2.6) and (2.7) via feasible Lasso. Therefore, the model selection properties of feasible Lasso as well as properties of least squares estimates for m and g based on the selected model play an important role in the derivation of the main result. The purpose of this appendix is to describe these properties.

Note that each of the regression models (2.6)-(2.7) obeys the following conditions.

Condition ASM: Approximate Sparse Model. *We observe $\omega_i = (\tilde{y}_i, \tilde{z}_i)$, $i = 1, \ldots, n$, where $\{\omega_i\}_{i=1}^{\infty}$ are i.n.i.d. vectors on the probability space $(\Omega, \mathcal{F}, \mathrm{P}_n)$, and we have that $\tilde{x}_i = P(\tilde{z}_i)$ for $1 \leqslant i \leqslant n$, where $P(z_i)$ is a p-dimensional dictionary of transformations of z_i, which may depend on n. These vectors that obey the following approximately sparse regression model for each n:*

$$\tilde{y}_i = f(\tilde{z}_i) + \epsilon_i = \tilde{x}_i'\beta_0 + r_i + \epsilon_i, \ \ \mathrm{E}[\epsilon_i \mid \tilde{x}_i] = 0, \ \ \bar{\mathrm{E}}[\epsilon_i^2] = \sigma^2, \ \ \|\beta_0\|_0 \leqslant s, \ \bar{\mathrm{E}}[r_i^2] \lesssim \sigma^2 s/n.$$

Let $\widehat{T}$ denote the model selected by the feasible Lasso estimator $\widehat{\beta}$:

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

The Post-Lasso estimator $\widetilde{\beta}$ is ordinary least squares applied to the data after removing the regressors that were not selected by the feasible Lasso:

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2] \ : \ \beta_j = 0 \text{ for each } j \notin \widehat{T}. \tag{B.1}$$

The following conditions are imposed to deal with non-Gaussian, heteroscedastic errors.

Condition RF: Reduced Form. *In addition to ASM, we have*

(i) $\log^3 p/n \to 0$ *and* $s\log(p \vee n)/n \to 0$,

(ii) $\bar{\mathrm{E}}[\tilde{y}_i^2] + \max_{1 \leqslant j \leqslant p}\{\bar{\mathrm{E}}[\tilde{x}_{ij}^2\tilde{y}_i^2] + \bar{\mathrm{E}}[|\tilde{x}_{ij}^3\epsilon_i^3|] + 1/\bar{\mathrm{E}}[\tilde{x}_{ij}^2\epsilon_i^2]\} \lesssim 1$,

(iii) $\max\limits_{1 \leqslant j \leqslant p}\{|(\mathbb{E}_n - \bar{\mathrm{E}})[\tilde{x}_{ij}^2\epsilon_i^2]| + |(\mathbb{E}_n - \bar{\mathrm{E}})[\tilde{x}_{ij}^2\tilde{y}_i^2]|\} + \max\limits_{1 \leqslant i \leqslant n}\|\tilde{x}_i\|_\infty^2 \dfrac{s\log(n \vee p)}{n} = o_P(1)$.

The main auxiliary result that we use in proving the main result is as follows.

Lemma 1. (Model Selection Properties of Lasso and Properties of Post-Lasso). *Let* $\{\mathrm{P}_n\}$ *be a sequence of data-generating processes. Suppose that conditions ASM and RF hold, and that Condition SE* (P_n) *holds for* $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$. *Consider a feasible Lasso estimator with penalty level and loadings specified as in Section 3.3.*

(i) Then the data-dependent model $\widehat{T}$ *selected by a feasible Lasso estimator satisfies with probability approaching 1:*

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

(ii) The Post-Lasso estimator obeys

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

Lemma 1 was derived in Belloni *et al*. (2012) for Iterated Lasso and by Belloni *et al*. (2010) for Square-root Lasso. These analyses build on the rate analysis of infeasible Lasso by Bickel *et al*. (2009) and on sparsity analysis and rate analysis of Post-Lasso by Belloni and Chernozhukov (2013).

APPENDIX C. PROOF OF THEOREM 1

The proof proceeds under given sequence of probability measures $\{\mathrm{P}_n\}$, as $n \to \infty$.

Let $Y = [y_1,\ldots,y_n]'$, $X = [x_1,\ldots,x_n]'$, $D = [d_1,\ldots,d_n]'$, $V = [v_1,\ldots,v_n]'$, $\zeta = [\zeta_1,\ldots,\zeta_n]'$, $m = [m_1,\ldots,m_n]'$, $R_m = [r_{m1},\ldots,r_{mn}]'$, $g = [g_1,\ldots,g_n]'$, $R_g = [r_{g1},\ldots,r_{gn}]'$, and so on. For $A \subset \{1,\ldots,p\}$, let $X[A] = \{X_j, j \in A\}$, where $\{X_j, j = 1,\ldots,p\}$ 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 span$[X[A]]$, and let $\mathcal{M}_A = \mathrm{I}_n - \mathcal{P}_A$ be the projection onto the subspace that is orthogonal to span$[X[A]]$. For a vector $W \in \mathbb{R}^n$, let

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

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

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

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

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

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 inequality, and because $\mathrm{E}[V'V/n]$ is bounded away from zero and from above uniformly in n by Condition SM, we have $ii^{-1} = \mathrm{E}[V'V/n]^{-1} + o_P(1)$.

By Condition SM we have $\sigma_n^2 = \bar{\mathrm{E}}[v_i^2]^{-1}\bar{\mathrm{E}}[\zeta_i^2 v_i^2]\bar{\mathrm{E}}[v_i^2]^{-1}$ is bounded away from zero and from above, uniformly in n. Hence

$$Z_n = \sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) = n^{-1/2}\sum_{i=1}^{n} z_{i,n} + o_P(1),$$

where $z_{i,n} := \sigma_n^{-1}\bar{\mathrm{E}}[v_i^2]^{-1}v_i\zeta_i$ are i.n.i.d. with mean zero. For $\delta > 0$ such that $4+2\delta \leqslant q$

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

by Condition SM. This condition verifies the Lyapunov condition and thus application of the Lyapunov CLT for i.n.i.d. triangular arrays implies that $Z_n \rightsquigarrow N(0,1)$.

Step 2. (Behavior of i.) Decompose, using $D = m+V$,

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

First, by Step 5 and 6 below we have

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

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

Second, using that $m = X\beta_{m0}+R_m$ and $m'\mathcal{M}_{\hat{I}}\zeta = R_m'\zeta - (\tilde{\beta}_m(\widehat{I})-\beta_{m0})'X'\zeta$, conclude

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

This follows since $|R_m'\zeta/\sqrt{n}| \lesssim_P \sqrt{R_m'R_m/n} \lesssim_P \sqrt{s/n}$, holding by Chebyshev inequality and Conditions SM and ASTE(iii), and

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

The latter bound follows by (a)

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

holding by Step 5 and by $\widehat{s} \lesssim_P s$ implied by Lemma 1, and (b) by $\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}$ holding by Step 4 under Condition SM.

Third, using similar reasoning, decomposition $g = X\beta_{g0}+R_g$, and Steps 4 and 6, conclude

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

Fourth, we have

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

since by Step 4 below $\|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}$, and

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

The latter bound follows from $\widehat{s} \lesssim_P s$, holding by Lemma 1, so that $\phi_{\min}^{-1}(\widehat{s}) \lesssim_P 1$ by Condition SE, and from $\|X'V/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}$ holding by Step 4.

Step 3. (Behaviour of ii.) Decompose

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

Then $|ii_a| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by Step 5, $|ii_b| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ 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: Bounds on $\|X'\zeta/\sqrt{n}\|_\infty$ and $\|X'V/\sqrt{n}\|_\infty$) Here we show that

$$\text{(a) } \|X'\zeta/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)} \text{ and (b)} \|X'V/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log(p\vee n)}.$$

To show (a), we use Lemma 3 stated in Appendix G on the tail bound for self-normalized deviations to deduce the bound. Indeed, we have that wp $\to 1$ for some $\ell_n \to \infty$ but so slowly that $1/\gamma = \ell_n \lesssim \log n$, with probability $1-o(1)$

$$\max_{1\leqslant j\leqslant p}\left|\frac{n^{-1/2}\sum_{i=1}^n x_{ij}\zeta_i}{\sqrt{\mathbb{E}_n[x_{ij}^2\zeta_i^2]}}\right| \leqslant \Phi^{-1}\left(1-\frac{1}{2\ell_n p}\right) \lesssim \sqrt{2\log(2\ell_n p)} \lesssim \sqrt{\log(p\vee n)}. \tag{C.3}$$

By Lemma 3 the first inequality in (C.3) holds, provided that for all n sufficiently large the following holds,

$$\Phi^{-1}\left(1-\frac{1}{2\ell_n p}\right)\leqslant \frac{n^{1/6}}{\ell_n}\min_{1\leqslant j\leqslant p} M_j^2-1,\ \ M_j:=\frac{\bar{\mathrm{E}}[x_{ij}^2\zeta_i^2]^{1/2}}{\bar{\mathrm{E}}[|x_{ij}^3||\zeta_i^3|]^{1/3}}.$$

Since we can choose ℓ_n to grow as slowly as needed, a sufficient condition for this are the conditions: $\log p = o(n^{1/3})$ and $\min_{1\leqslant j\leqslant p} M_j \gtrsim 1$, which both hold by Condition SM. Finally,

$$\max_{1\leqslant j\leqslant p}\mathbb{E}_n[x_{ij}^2\zeta_i^2]\lesssim_{\mathrm{P}} 1, \tag{C.4}$$

by Condition SM. Therefore (a) follows from the bounds (C.3) and (C.4). Claim (b) follows similarly.

Step 5. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}}m\|$ and related quantities.) This step shows that

$$\text{(a) } \|\mathcal{M}_{\widehat{I}}m/\sqrt{n}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n} \text{ and (b) } \|\tilde{\beta}_m(\widehat{I})-\beta_{m0}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n}.$$

Observe that

$$\sqrt{[s\log(p\vee n)]/n}\underset{(1)}{\gtrsim_P}\|\mathcal{M}_{\widehat{I}_1}m/\sqrt{n}\|\underset{(2)}{\gtrsim_P}\|\mathcal{M}_{\widehat{I}}m/\sqrt{n}\|$$

where inequality (1) holds since by Lemma 1 $\|\mathcal{M}_{\widehat{I}_1}m/\sqrt{n}\|\leqslant\|(X\tilde{\beta}_D(\widehat{I}_1)-m)/\sqrt{n}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n}$, and (2) holds by $\widehat{I}_1\subseteq\widehat{I}$ by construction. This shows claim (a). To show claim (b) note that

$$\|\mathcal{M}_{\widehat{I}}m/\sqrt{n}\|\underset{(3)}{\geqslant}|\|X(\tilde{\beta}_m(\widehat{I})-\beta_{m0})/\sqrt{n}\|-\|R_m/\sqrt{n}\||$$

where (3) holds by the triangle inequality. Since $\|R_m/\sqrt{n}\|\lesssim_P\sqrt{s/n}$ by Chebyshev and Condition ASTE(iii), conclude that

$$\begin{aligned}\sqrt{[s\log(p\vee n)]/n} &\gtrsim_P \|X(\tilde{\beta}_m(\widehat{I})-\beta_{m0})/\sqrt{n}\|\\ &\geqslant \sqrt{\phi_{\min}(\widehat{s}+s)}\|\tilde{\beta}_m(\widehat{I})-\beta_{m0}\|\gtrsim_P\|\tilde{\beta}_m(\widehat{I})-\beta_{m0}\|,\end{aligned}$$

since $\widehat{s}\lesssim_P s$ by Lemma 1 so that $1/\phi_{\min}(\widehat{s}+s)\lesssim_P 1$ by condition SE. This shows claim (b).

Step 6. (Auxiliary: Bound on $\|\mathcal{M}_{\widehat{I}}g\|$ and related quantities.) This step shows that

$$\text{(a) } \|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n} \text{ and (b) } \|\tilde{\beta}_g(\widehat{I})-\beta_{g0}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n}.$$

Observe that

$$\begin{aligned}\sqrt{[s\log(p\vee n)]/n} &\underset{(1)}{\gtrsim_P} \|\mathcal{M}_{\widehat{I}_2}(\alpha_0 m+g)/\sqrt{n}\|\underset{(2)}{\gtrsim_P}\|\mathcal{M}_{\widehat{I}}(\alpha_0 m+g)/\sqrt{n}\|\\ &\underset{(3)}{\gtrsim_P} |\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\|-\|\mathcal{M}_{\widehat{I}}\alpha_0 m/\sqrt{n}\||\end{aligned}$$

where inequality (1) holds since by Lemma 1 $\|\mathcal{M}_{\widehat{I}_2}(\alpha_0 m+g)/\sqrt{n}\|\leqslant\|(X\tilde{\beta}_Y(\widehat{I}_2)-\alpha_0 m-g)/\sqrt{n}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n}$, (2) holds by $\widehat{I}_2\subseteq\widehat{I}$, and (3) by the triangle inequality. Since $|\alpha_0|$ is bounded uniformly in n by assumption, by Step 5, $\|\mathcal{M}_{\widehat{I}}\alpha_0 m/\sqrt{n}\|\lesssim_P\sqrt{[s\log(p\vee n)]/n}$. Hence claim (a) follows by the triangle inequality:

$$\sqrt{[s\log(p\vee n)]/n}\gtrsim_P\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\|$$

To show claim (b) we note that $\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\|\geqslant|\|X(\tilde{\beta}_g(\widehat{I})-\beta_{g0})/\sqrt{n}\|-\|R_g/\sqrt{n}\||$, where $\|R_g/\sqrt{n}\|\lesssim_P\sqrt{s/n}$ by Condition ASTE(iii). Then conclude similarly to Step 5 that

$$\sqrt{[s\log(p\vee n)]/n}\gtrsim_P\|X(\tilde{\beta}_g(\widehat{I})-\beta_{g0})/\sqrt{n}\|\geqslant\sqrt{\phi_{\min}(\widehat{s}+s)}\|\tilde{\beta}_g(\widehat{I})-\beta_{g0}\|\gtrsim_P\|\tilde{\beta}_g(\widehat{I})-\beta_{g0}\|.$$

Step 7. (Variance Estimation.) Since $\widehat{s}\lesssim_P s=o(n)$, $(n-\widehat{s}-1)/n=o_P(1)$, and since $\bar{\mathrm{E}}[v_i^2\zeta_i^2]$ and $\bar{\mathrm{E}}[v_i^2]$ are bounded away from zero and from above uniformly in n by Condition SM, it suffices to show that $\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2]-\bar{\mathrm{E}}[v_i^2\zeta_i^2]\to_P 0$, $\mathbb{E}_n[\widehat{v}_i^2]-\bar{\mathrm{E}}[v_i^2]\to_P 0$, The second relation was shown in Step 3, so it remains to show the first relation.

Let $\tilde{v}_i=v_i+r_{mi}$ and $\tilde{\zeta}_i=\zeta_i+r_{gi}$. Recall that by Condition ASTE(v) we have $\bar{\mathrm{E}}[\tilde{v}_i^2\tilde{\zeta}_i^2]-\bar{\mathrm{E}}[v_i^2\zeta_i^2]\to 0$, and $\mathbb{E}_n[\tilde{v}_i^2\tilde{\zeta}_i^2]-\bar{\mathrm{E}}[\tilde{v}_i^2\tilde{\zeta}_i^2]\to_P 0$ by Vonbahr–Esseen's inequality in von Bahr and Esseen (1965) since $\bar{\mathrm{E}}[|\tilde{v}_i\tilde{\zeta}_i|^{2+\delta}]\leqslant(\bar{\mathrm{E}}[|\tilde{v}_i|^{4+2\delta}]\bar{\mathrm{E}}[|\tilde{\zeta}_i|^{4+2\delta}])^{1/2}$ is uniformly bounded for $4+2\delta\leqslant q$. Thus it suffices to show that $\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2]-\mathbb{E}_n[\tilde{v}_i^2\tilde{\zeta}_i^2]\to_P 0$.

By the triangle inequality

$$|\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2-\tilde{v}_i^2\tilde{\zeta}_i^2]|\leqslant\underbrace{|\mathbb{E}_n[(\widehat{v}_i^2-\tilde{v}_i^2)\tilde{\zeta}_i^2]|}_{=:iv}+\underbrace{|\mathbb{E}_n[\widehat{v}_i^2(\widehat{\zeta}_i^2-\tilde{\zeta}_i^2)]|}_{=:iii}.$$

Then, expanding $\widehat{\zeta}_i^2-\tilde{\zeta}_i^2$ we have

$$\begin{aligned} iii &\leqslant 2\mathbb{E}_n[\{d_i(\alpha_0-\check{\alpha})\}^2\widehat{v}_i^2]+2\mathbb{E}_n[\{x_i'(\check{\beta}-\beta_{g0})\}^2\widehat{v}_i^2]+|2\mathbb{E}_n[\tilde{\zeta}_i d_i(\alpha_0-\check{\alpha})\widehat{v}_i^2]|+|2\mathbb{E}_n[\tilde{\zeta}_i x_i'(\check{\beta}-\beta_{g0})\widehat{v}_i^2]|\\ &=: iii_a+iii_b+iii_c+iii_d=o_P(1)\end{aligned}$$

where the last bound follows by the relations derived below.

First, we note

$$iii_a \leqslant 2\max_{i\leqslant n} d_i^2|\alpha_0-\check{\alpha}|^2\mathbb{E}_n[\widehat{v}_i^2]\lesssim_P n^{(2/q)-1}=o(1) \tag{C.5}$$

$$iii_c \leqslant 2\max_{i\leqslant n}\{|\tilde{\zeta}_i||d_i|\}\mathbb{E}_n[\widehat{v}_i^2]|\alpha_0-\check{\alpha}|\lesssim_P n^{(2/q)-(1/2)}=o(1) \tag{C.6}$$

which holds by the following argument. Condition SM assumes that $\mathrm{E}[|d_i|^q]$ which in turn implies that $\mathrm{E}[\max_{i\leqslant n} d_i^2]\lesssim n^{2/q}$. Similarly Condition ASTE implies that $\mathrm{E}[\max_{i\leqslant n}\tilde{\zeta}_i^2]\lesssim n^{2/q}$ and $\mathrm{E}[\max_{i\leqslant n}\tilde{v}_i^2]\lesssim n^{2/q}$. Thus by Markov inequality

$$\max_{i\leqslant n}|d_i|+|\tilde{\zeta}_i|+|\tilde{v}_i|\lesssim_P n^{1/q}. \tag{C.7}$$

Moreover, $\mathbb{E}_n[\widehat{v}_i^2]\lesssim_P 1$ and $|\check{\alpha}-\alpha_0|\lesssim_P n^{-1/2}$ by the previous steps. These bounds and $q>4$ imposed in Condition SM imply (C.5)-(C.6).

Next we bound,

$$iii_d \leqslant 2\max_{i\leqslant n}|\tilde{\zeta}_i|\max_{i\leqslant n}|x_i'(\check{\beta}-\beta_{g0})|\mathbb{E}_n[\widehat{v}_i^2]\lesssim_P n^{1/q}\max_{i\leqslant n}\|x_i\|_\infty\sqrt{\frac{s}{\sqrt{n}}\frac{s\log(p\vee n)}{\sqrt{n}}}=o_P(1). \tag{C.8}$$

To establish this we use (C.7) and that for $\widehat{T}_g=\text{support}(\beta_{g0})\cup\widehat{I}$, $\max_{i\leqslant n}\{x_i'(\check{\beta}-\beta_{g0})\}^2\leqslant\max_{i\leqslant n}\|x_{i\widehat{T}_g}\|^2\|\check{\beta}-\beta_{g0}\|^2$, where $\max_{i\leqslant n}\|x_{i\widehat{T}_g}\|^2\leqslant|\widehat{T}_g|\max_{i\leqslant n}\|x_i\|_\infty^2\lesssim_P s\max_{i\leqslant n}\|x_i\|_\infty^2$ by the sparsity assumption in ASTE and the sparsity bound in Lemma 1. Since $\check{\beta}[\widehat{I}]=(X[\widehat{I}]'X[\widehat{I}])^{-}X[\widehat{I}]'(\zeta+g-(\check{\alpha}-\alpha_0)D)$,

$$\|\check{\beta}-\beta_{g0}\|\leqslant\|\tilde{\beta}_g(\widehat{I})-\beta_{g0}\|+\|\tilde{\beta}_\zeta(\widehat{I})\|+|\check{\alpha}-\alpha_0|\cdot\|\tilde{\beta}_D(\widehat{I})\|\lesssim_P\sqrt{s\log(p\vee n)/n}.$$

The last inequality follows by Step 6(b), by $\|\tilde{\beta}_\zeta(\widehat{I})\|\leqslant\sqrt{\widehat{s}}\phi_{\min}^{-1}(\widehat{s})\|X'\zeta/n\|_\infty\lesssim_P\sqrt{s\log(p\vee n)/n}$ holding by Condition SE and by $\widehat{s}\lesssim_P s$ from Lemma 1, and by Step 4, $|\check{\alpha}-\alpha_0|\lesssim_P 1/\sqrt{n}$ by Step 1, and $\|\tilde{\beta}_D(\widehat{I})\|\leqslant\phi_{\min}^{-1}(\widehat{s})\sqrt{\widehat{s}}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}d_i]|\leqslant\phi_{\min}^{-1}(\widehat{s})\sqrt{\widehat{s}}\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[x_{ij}^2d_i^2]}\lesssim_P\sqrt{s}$ by Condition SE, $\widehat{s}\lesssim_P s$ by the sparsity bound in Lemma 1, and Condition SM.

The final conclusion in (C.8) then follows by condition ASTE (iv) and (v).

Next, using the relations above and condition ASTE (iv) and (v), we also conclude that

$$iii_b \leqslant 2\max_{i\leqslant n}\{x_i'(\check{\beta}-\beta_{g0})\}^2\mathbb{E}_n[\widehat{v}_i^2]\lesssim_P\max_{i\leqslant n}\|x_i\|_\infty^2\frac{s}{\sqrt{n}}\frac{s\log(p\vee n)}{\sqrt{n}}=o_P(1).$$

Finally, the argument for $iv=o_P(1)$ follows similarly to the argument for $iii=o_P(1)$ and the result follows. ‖

APPENDIX D. PROOF OF COROLLARY 1

Let $\mathbf{P}_n$ be a collection of probability measures P for which conditions ASTE (P), SM (P), SE (P), and R (P) hold for the given n. Consider any sequence $\{\mathrm{P}_n\}$, with index $n\in\{1,2,...\}$, with $\mathrm{P}_n\in\mathbf{P}_n$ for each $n\in\{1,2,...\}$. By Theorem 1 we have that, for $c=\Phi^{-1}(1-\gamma/2)$, $\lim_{n\to\infty}\mathrm{P}_n\left(\alpha_0\in[\check{\alpha}\pm c\widehat{\sigma}_n/\sqrt{n}]\right)=\Phi(c)-\Phi(-c)=1-\gamma$. This means that for every further subsequence $\{\mathrm{P}_{n_k}\}$ with $\mathrm{P}_{n_k}\in\mathbf{P}_{n_k}$ for each $k\in\{1,2,...\}$

$$\lim_{k\to\infty}\mathrm{P}_{n_k}\left(\alpha_0\in[\check{\alpha}\pm c\widehat{\sigma}_{n_k}/\sqrt{n_k}]\right)=1-\gamma. \tag{D.9}$$

Suppose that $\limsup_{n\to\infty}\sup_{\mathrm{P}\in\mathbf{P}_n}\left|\mathrm{P}\left(\alpha_0\in[\check{\alpha}\pm c\widehat{\sigma}_n/\sqrt{n}]\right)-(1-\gamma)\right|>0$. Hence there is a subsequence $\{\mathrm{P}_{n_k}\}$ with $\mathrm{P}_{n_k}\in\mathbf{P}_{n_k}$ for each $k\in\{1,2,...\}$ such that: $\lim_{k\to\infty}\mathrm{P}_{n_k}\left(\alpha_0\in[\check{\alpha}\pm c\widehat{\sigma}_{n_k}/\sqrt{n_k}]\right)\neq 1-\gamma$. This gives a contradiction to (D.9). It then follows that

$$\lim_{n\to\infty}\sup_{\mathrm{P}\in\mathbf{P}_n}|\mathrm{P}\left(\alpha_0\in[\check{\alpha}\pm c(1-\xi)\widehat{\sigma}_n/\sqrt{n}]\right)-(1-\xi)|=0.$$

The claim then follows from this since $\mathbf{P}\subseteq\mathbf{P}_n$ for all $n\geqslant n_0$. ‖

APPENDIX E. PROOF OF THEOREM 2

We use the same notation as in Theorem 1. Using that notation the approximations bounds stated in Condition HLMS are equivalent to $\|\mathcal{M}_{\widehat{I}}g\|\leqslant\delta_n n^{1/4}$ and $\|\mathcal{M}_{\widehat{I}}m\|\leqslant\delta_n n^{1/4}$.

Step 1. It follows the same reasoning as Step 1 in the proof of Theorem 1.

Step 2. (Behavior of i.) Decompose, using $D=m+V$

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

First, by Condition HLMS we have $\|\mathcal{M}_{\hat I}g\|=o_P(n^{1/4})$ and $\|\mathcal{M}_{\hat I}m\|=o_P(n^{1/4})$. Therefore

$$|i_a|=|m'\mathcal{M}_{\hat I}g/\sqrt{n}|\leqslant\sqrt{n}\|\mathcal{M}_{\hat I}g/\sqrt{n}\|\|\mathcal{M}_{\hat I}m/\sqrt{n}\|\lesssim_P o(1).$$

Second, using that $m=X\beta_{m0}+R_m$ and $m'\mathcal{M}_{\hat I}\zeta=R_m'\zeta-(\tilde\beta_m(\widehat{I})-\beta_{m0})'X'\zeta$, we have

$$\begin{aligned}|i_b|&\leqslant|R_m'\zeta/\sqrt{n}|+|(\tilde\beta_m(\widehat{I})-\beta_{m0})'X'\zeta/\sqrt{n}|\leqslant|R_m'\zeta/\sqrt{n}|+\|\tilde\beta_m(\widehat{I})-\beta_{m0}\|_1\|X'\zeta/\sqrt{n}\|_\infty\\&\lesssim_P\sqrt{s/n}+\sqrt{s}\,\{o(n^{-1/4})+\sqrt{s/n}\}\sqrt{\log(p\vee n)}=o(1).\end{aligned}$$

This follows because $|R_m'\zeta/\sqrt{n}|\lesssim_P\sqrt{R_m'R_m/n}\lesssim_P\sqrt{s/n}$, by Chebyshev inequality and Conditions SM and ASTE(iii),

$$\|\tilde\beta_m(\widehat{I})-\beta_{m0}\|_1\leqslant\sqrt{\widehat{s}+s}\|\tilde\beta_m(\widehat{I})-\beta_{m0}\|\lesssim_P\sqrt{s}\,\{o(n^{-1/4})+\sqrt{s/n}\},$$

by Step 4 and $\widehat{s}=|\widehat{I}|\lesssim_P s$ by Condition HLMS, and $\|X'\zeta/\sqrt{n}\|_\infty\lesssim_P\sqrt{\log(p\vee n)}$ holding by Step 4 in the proof of Theorem 1.

Third, using similar reasoning and the decomposition $g=X\beta_{g0}+R_g$ conclude

$$\begin{aligned}|i_c|&\leqslant|R_g'V/\sqrt{n}|+|(\tilde\beta_g(\widehat{I})-\beta_{g0})'X'V/\sqrt{n}|\\&\lesssim_P\sqrt{s/n}+\sqrt{s}\,\{o(n^{-1/4})+\sqrt{s/n}\}\,\sqrt{\log(p\vee n)}=o_P(1).\end{aligned}$$

Fourth, we have

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

since $\|X'\zeta/\sqrt{n}\|_\infty\lesssim_P\sqrt{\log(p\vee n)}$ by Step 4 of the proof of Theorem 1, and

$$\begin{aligned}\|\tilde\beta_V(\widehat{I})\|_1&\leqslant\sqrt{\widehat{s}}\|\tilde\beta_V(\widehat{I})\|\leqslant\sqrt{\widehat{s}}\|(X[\widehat{I}]'X[\widehat{I}]/n)^{-1}X[\widehat{I}]'V/n\|\\&\leqslant\sqrt{\widehat{s}}\phi_{\min}^{-1}(\widehat{s})\sqrt{\widehat{s}}\|X'V/\sqrt{n}\|_\infty/\sqrt{n}\lesssim_P s\sqrt{[\log(p\vee n)]/n}.\end{aligned}$$

The latter bound follows from $\widehat{s}\lesssim_P s$ by condition HLMS so that $\phi_{\min}^{-1}(\widehat{s})\lesssim_P 1$ by condition SE, and again invoking Step 4 of the proof of Theorem 1 to establish $\|X'V/\sqrt{n}\|_\infty\lesssim_P\sqrt{\log(p\vee n)}$.

Step 3. (Behaviour of ii.) Decompose

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

Then $|ii_a|\lesssim_P o(n^{1/2})/n=o_P(n^{-1/2})$ by condition HLMS, $|ii_b|=o(n^{-1/2})$ 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: Bounds on $\|\tilde\beta_m(\widehat{I})-\beta_{m0}\|$ and $\|\tilde\beta_g(\widehat{I})-\beta_{g0}\|$.) To establish a bound on $\|\tilde\beta_g(\widehat{I})-\beta_{g0}\|$ note that $\|\mathcal{M}_{\hat I}g/\sqrt{n}\|\geqslant|\,\|X(\tilde\beta_g(\widehat{I})-\beta_{g0})/\sqrt{n}\|-\|R_g/\sqrt{n}\|\,|$, where $\|R_g/\sqrt{n}\|\lesssim_P\sqrt{s/n}$ holds by Chebyshev inequality and Condition ASTE(iii). Moreover, by Condition HLMS we have $\|\mathcal{M}_{\hat I}g/\sqrt{n}\|=o_P(n^{-1/4})$ and $\widehat{s}=|\widehat{I}|\lesssim_P s$. Thus

$$o(n^{-1/4})+\sqrt{s/n}\gtrsim_P\|X(\tilde\beta_g(\widehat{I})-\beta_{g0})/\sqrt{n}\|\geqslant\sqrt{\phi_{\min}(s+\widehat{s})}\|\tilde\beta_g(\widehat{I})-\beta_{g0}\|\gtrsim_P\|\tilde\beta_g(\widehat{I})-\beta_{g0}\|$$

since $\sqrt{\phi_{\min}(s+\widehat{s})}\gtrsim_P 1$ by Condition SE. The same logic yields $\|\tilde\beta_m(\widehat{I})-\beta_{m0}\|\lesssim_P\sqrt{s/n}+o(n^{-1/4})$.

Step 5. (Variance Estimation.) It follows similarly to Step 7 in the proof of Theorem 1 but using Condition HLMS instead of Lemma 1.

Step 6. (Uniformity Properties) The proof is similar to the proof of Corollary 1. ‖

APPENDIX F. PROOF OF THEOREM 3

The two results have identical structure and have nearly the same proof, and so we present the proof only for the case of considering the ATE parameter value $\alpha_0=\mathrm{E}[g(1,z_i)-g(0,z_i)]$.

In the proof $a\lesssim b$ means that $a\leqslant Ab$, where the constant A depends on the constants in Condition HT only, but not on n once $n\geqslant n_0=\min\{j:\delta_j\leqslant 1/2\}$, and not on $\mathrm{P}\in\mathbf{P}_n$. For the proof of claims (1) and (2) we consider a sequence P_n in $\mathbf{P}_n$, but for simplicity, we write P throughout the proof, omitting the index n. Since the argument is asymptotic, we can just assume that $n\geqslant n_0$ in what follows.

Step 1. In this step we establish claim (1).

(a) We begin with a preliminary observation. Define, for $t=(t_1,t_2,t_3)\in\mathbb{R}^2\times(0,1)$,

$$\psi(y,d,t)=\frac{d(y-t_2)}{t_3}-\frac{(1-d)(y-t_1)}{1-t_3}-t_2-t_1.$$

The derivatives of this function with respect to t obey for all $k=(k_j)_{j=1}^3\in\mathbb{N}^3:0\leqslant|k|\leqslant 3$,

$$|\partial_t^k\psi(y,d,t)|\leqslant L,\ \ \forall(y,d,t):|y|\leqslant C,|t_1|\leqslant C,|t_2|\leqslant C,c'/2\leqslant|t_3|\leqslant 1-c'/2, \tag{F.10}$$

where L depends only on c' and C, $|k|=\sum_{j=1}^3 k_j$, and $\partial_t^k:=\partial_{t_1}^{k_1}\partial_{t_2}^{k_2}\partial_{t_3}^{k_3}$.

(b). Let

$$\widehat{h}(z_i):=(\widehat{g}(0,z_i),\widehat{g}(1,z_i),\widehat{m}(z_i))',\ \ h_0(z_i):=(g(0,z_i),g(1,z_i),m(z_i))',$$
$$f_{\widehat{h}}(y_i,d_i,z_i):=\psi(y_i,d_i,\widehat{h}(z_i)),\ \ f_{h_0}(y_i,d_i,z_i):=\psi(y_i,d_i,h_0(z_i)).$$

We observe that with probability no less than $1-\Delta_n$,

$$\widehat{g}(0,\cdot)\in\mathcal{G}_0,\ \ \widehat{g}(1,\cdot)\in\mathcal{G}_1 \text{ and } \widehat{m}\in\mathcal{M},$$

$$\mathcal{G}_d:=\{z\mapsto x'\beta:\|\beta\|_0\leqslant sC,\|x_i'\beta-g(d,z_i)\|_{\mathrm{P},2}\lesssim\delta_n n^{-1/4},\|x_i'\beta-g(d,z_i)\|_{\mathrm{P},\infty}\lesssim\delta_n\},$$

$$\mathcal{M}:=\{z\mapsto\Lambda(x'\beta):\|\beta\|_0\leqslant sC,\|\Lambda(x_i'\beta)-m(z_i)\|_{\mathrm{P},2}\lesssim\delta_n n^{-1/4},\|\Lambda(x_i'\beta)-m(z_i)\|_{\mathrm{P},\infty}\lesssim\delta_n\}.$$

To see this note, that under assumption HT (P), under condition (i)-(ii), under the event occurring under condition (ii) of that assumption: for $n\geqslant n_0=\min\{j:\delta_j\leqslant 1/2\}$:

$$\|\tilde{x}_i'\beta-g(d_i,z_i)\|_{\mathrm{P},2}\leqslant\|\tilde{x}_i'(\beta-\beta_{g0})\|_{\mathrm{P},2}+\|r_{gi}\|_{\mathrm{P},2}\leqslant 2\|\tilde{x}_i'(\beta-\beta_{g0})\|_{\mathbb{P}_n,2}+\|r_{gi}\|_{\mathrm{P},2}\leqslant 4\delta_n n^{-1/4},$$

$$\|\tilde{x}_i'\beta-g(d_i,z_i)\|_{\mathrm{P},\infty}\leqslant\|\tilde{x}_i'(\beta-\beta_{g0})\|_{\mathrm{P},\infty}+\|r_{gi}\|_{\mathrm{P},\infty}\leqslant K_n\|\beta-\beta_{0g}\|_1+\delta_n\leqslant 2\delta_n,$$

for $\beta=\widehat{\beta}_g$, with evaluation after computing the norms, and noting that for any β

$$\|x_i'\beta-g(1,z_i)\|_{\mathrm{P},2}\vee\|x_i'\beta-g(0,z_i)\|_{\mathrm{P},2}\lesssim\|\tilde{x}_i'\beta-g(d_i,z_i)\|_{\mathrm{P},2}$$

under condition (iii). Furthermore, for $n\geqslant n_0=\min\{j:\delta_j\leqslant 1/2\}$:

$$\|\Lambda(x_i'\beta)-m(z_i)\|_{\mathrm{P},2}\leqslant\|\Lambda(x_i'\beta)-\Lambda(x_i'\beta_{m0})\|_{\mathrm{P},2}+\|r_{mi}\|_{\mathrm{P},2}$$

$$\lesssim\|\partial\Lambda\|_\infty\|\tilde{x}_i'(\beta-\beta_{m0})\|_{\mathrm{P},2}+\|r_{mi}\|_{\mathrm{P},2}\lesssim\|\partial\Lambda\|_\infty\|\tilde{x}_i'(\beta-\beta_{m0})\|_{\mathbb{P}_n,2}+\|r_{mi}\|_{\mathrm{P},2}\lesssim\delta_n n^{-1/4}$$

$$\|\Lambda(x_i'\beta)-m(z_i)\|_{\mathrm{P},\infty}\leqslant\|\partial\Lambda\|_\infty\|\tilde{x}_i'(\beta-\beta_{g0})\|_{\mathrm{P},\infty}+\|r_{mi}\|_{\mathrm{P},\infty}\lesssim K_n\|\beta-\beta_{m0}\|_1+\delta_n\leqslant 2\delta_n,$$

for $\beta=\widehat{\beta}_{m0}$, with evaluation after computing the norms.

Hence with probability at least $1-\Delta_n$,

$$\widehat{h}\in\mathcal{H}_n:=\{h=(\bar{g}(0,z),\bar{g}(1,z),\bar{m}(z))\in\mathcal{G}_0\times\mathcal{G}_1\times\mathcal{M}\}.$$

(c) We have that

$$\alpha_0=\mathrm{E}[f_{h_0}] \text{ and } \check{\alpha}=\mathbb{E}_n[f_{\widehat{h}}],$$

so that

$$\sqrt{n}(\check{\alpha}-\alpha_0)=\underbrace{\mathbb{G}_n[f_{h_0}]}_{i}+\underbrace{(\mathbb{G}_n[f_h]-\mathbb{G}_n[f_{h_0}])}_{ii}+\underbrace{\sqrt{n}(\mathrm{E}[f_h-f_{h_0}])}_{iii},$$

with h evaluated at $h=\widehat{h}$. By the Lyapunov central limit theorem, $\sigma_n^{-1}i\rightsquigarrow N(0,1)$.

(d) Note that for $\Delta_i=h(z_i)-h_0(z_i)$, and Δ_i^k denoting $\Delta_{i1}^{k_1}\Delta_{i2}^{k_2}\Delta_{i3}^{k_3}$,

$$\begin{aligned} iii=&\sqrt{n}\sum_{|k|=1}\mathrm{E}[\partial_t^k\psi(y_i,d_i,h_0(z_i))\Delta_i^k]\\ &+\sqrt{n}\sum_{|k|=2}2^{-1}\mathrm{E}[\partial_t^k\psi(y_i,d_i,h_0(z_i))\Delta_i^k]\\ &+\sqrt{n}\sum_{|k|=3}\int_0^1 6^{-1}\mathrm{E}[\partial_t^k\psi(y_i,d_i,h_0(z_i)+\lambda\Delta_i)\Delta_i^k]d\lambda\ \ =:iii_a+iii_b+iii_c,\end{aligned}$$

(with h evaluated at $h=\widehat{h}$). By the law of iterated expectations and because

$$\mathrm{E}[\partial_t^k\psi(y_i,d_i,h_0(z_i))|z_i]=0\ \ \forall k\in\mathbb{N}^3:|k|=1,$$

we have that $iii_a=0$. Moreover, uniformly for any $h\in\mathcal{H}_n$ we have that

$$|iii_b|\lesssim\sqrt{n}\|h-h_0\|_{\mathrm{P},2}^2\lesssim\sqrt{n}(\delta_n n^{-1/4})^2\leqslant\delta_n^2,$$
$$|iii_c|\lesssim\sqrt{n}\|h-h_0\|_{\mathrm{P},2}^2\|h-h_0\|_{\mathrm{P},\infty}\lesssim\sqrt{n}(\delta_n n^{-1/4})^2\delta_n\leqslant\delta_n^3.$$

Since $\widehat{h}\in\mathcal{H}_n$ with probability $1-\Delta_n$, we have that once $n\geqslant n_0$, $\mathrm{P}(|iii|\lesssim\delta_n^2)\geqslant 1-\Delta_n$.

(e). Furthermore, we have that $|ii| \leqslant \sup_{h\in\mathcal{H}_n} |\mathbb{G}_n[f_h]-\mathbb{G}_n[f_{h_0}]|$.

The class of functions $\mathcal{G}_d$ for $d\in\{0,1\}$ is a union of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$. The class of functions $\mathcal{M}$ is a subset of a union of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$ (monotone transformation Λ preserve the VC-subgraph property). These classes are uniformly bounded and their entropies therefore satisfy

$$\log N(\varepsilon,\mathcal{M},\|\cdot\|_{\mathbb{P}_n,2})+\log N(\varepsilon,\mathcal{G}_0,\|\cdot\|_{\mathbb{P}_n,2})+\log N(\varepsilon,\mathcal{G}_1,\|\cdot\|_{\mathbb{P}_n,2})\lesssim s\log p+s\log(1/\varepsilon).$$

Finally, the class $\mathcal{F}_n=\{f_h-f_{h_0}:h\in\mathcal{H}_n\}$ is a Lipschitz transform of $\mathcal{H}_n$ with bounded Lipschitz coefficients and with a constant envelope. Therefore, we have that

$$\log N(\varepsilon,\mathcal{F}_n,\|\cdot\|_{\mathbb{P}_n,2})\lesssim s\log p+s\log(1/\varepsilon).$$

We shall invoke the following lemma derived in Belloni and Chernozhukov (2011a).

Lemma 2. (A Self-Normalized Maximal inequality). *Let $\mathcal{F}$ be a measurable function class on a sample space. Let $F=\sup_{f\in\mathcal{F}}|f|$, and suppose that there exist some constants $\omega_n>3$ and $\upsilon>1$, such hat*

$$\log N(\epsilon\|F\|_{\mathbb{P}_n,2},\mathcal{F},\|\cdot\|_{\mathbb{P}_n,2})\leqslant \upsilon m(\log(n\vee\omega_n)+\log(1/\epsilon)),\ 0<\epsilon<1.$$

Then for every $\delta\in(0,1/6)$ we have

$$\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)|\leqslant C_\upsilon\sqrt{2/\delta}\sqrt{m\log(n\vee\omega_n)}(\sup_{f\in\mathcal{F}}\|f\|_{\mathrm{P},2}\vee\sup_{f\in\mathcal{F}}\|f\|_{\mathbb{P}_n,2}),$$

with probability at least $1-\delta$ for some constant that C_υ.

Then by Lemma 2 together and some simple calculations, we have that

$$\begin{aligned}|ii| &\leqslant \sup_{f\in\mathcal{F}_n}|\mathbb{G}_n(f)|=O_{\mathrm{P}}(1)\sqrt{s\log(p\vee n)}(\sup_{f\in\mathcal{F}_n}\|f\|_{\mathbb{P}_n,2}\vee\sup_{f\in\mathcal{F}_n}\|f\|_{\mathrm{P},2})\\ &\leqslant O_{\mathrm{P}}(1)\sqrt{s\log(p\vee n)}(\sup_{h\in\mathcal{H}_n}\|h-h_0\|_{\mathbb{P}_n,2}\vee\sup_{h\in\mathcal{H}_n}\|h-h_0\|_{\mathrm{P},2})=o_{\mathrm{P}}(1).\end{aligned}$$

The last conclusion follows because $\sup_{h\in\mathcal{H}_n}\|h-h_0\|_{\mathrm{P},2}\lesssim\delta_n n^{-1/4}$ by definition of $\mathcal{H}_n$, and

$$\sup_{h\in\mathcal{H}_n}\|h-h_0\|_{\mathbb{P}_n,2}\quad\leqslant O_{\mathrm{P}}(1)\cdot\left(\sup_{h\in\mathcal{H}_n}\|h-h_0\|_{\mathrm{P},2}+\|r_{gi}\|_{\mathrm{P},2}+\|r_{mi}\|_{\mathrm{P},2}\right),$$

where the last conclusion follows from the same argument as in step (b) but in a reverse order, switching from empirical norms to population norms, using equivalence of norms over sparse sets imposed in condition (ii) , and also using an application of Markov inequality to argue that $\|r_{gi}\|_{\mathbb{P}_n,2}+\|r_{mi}\|_{\mathbb{P}_n,2}=O_{\mathrm{P}}(1)(\|r_{gi}\|_{\mathrm{P},2}+\|r_{mi}\|_{\mathrm{P},2})$.

Step 2. Claim (2) follows from consistency: $\widehat{\sigma}_n/\sigma_n=1+o_{\mathrm{P}}(1)$, which follows from $\widehat{\sigma}_n$ being a Lipschitz transform of $\widehat{h}$ with respect to $\|\cdot\|_{\mathbb{P}_n,2}$, once $\widehat{h}\in\mathcal{H}_n$ and the consistency of $\widehat{h}$ for h under $\|\cdot\|_{\mathbb{P}_n,2}$.

Step 3. Claim (3) is immediate from claims (1) and (2) by the way of contradiction as in the proof of Corollary 1. ||

APPENDIX G. MODERATE DEVIATIONS FOR A MAXIMUM OF SELF-NORMALIZED AVERAGES

Lemma 3. (Moderate Deviation Inequality for Maximum of a Vector). *Suppose that*

$$\mathcal{S}_j=\frac{\sum_{i=1}^n U_{ij}}{\sqrt{\sum_{i=1}^n U_{ij}^2}},$$

where U_{ij} are independent variables across i with mean zero. We have that

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathcal{S}_j|>\Phi^{-1}(1-\gamma/2p)\right)\leqslant\gamma\left(1+\frac{A}{\ell_n^3}\right),$$

where A is an absolute constant, provided for $\ell_n>0$

$$0\leqslant\Phi^{-1}(1-\gamma/(2p))\leqslant\frac{n^{1/6}}{\ell_n}\min_{1\leqslant j\leqslant p}M_j^2-1,\quad M_j:=\frac{\left(\frac{1}{n}\sum_{i=1}^n\mathrm{E}[U_{ij}^2]\right)^{1/2}}{\left(\frac{1}{n}\sum_{i=1}^n\mathrm{E}[|U_{ij}|^3]\right)^{1/3}}.$$

This results is essentially due to Jing *et al*. (2003). The proof of this result, given in Belloni *et al*. (2012), follows from a simple combination of union bounds with the bounds in Theorem 7.4 in de la Peña *et al*. (2009), which was originally derived by Jing *et al*. (2003).

Acknowledgments. We thank five anonymous referees, the editor of the Review, Stéphane Bonhomme, the co-editor of Econometrica, James Stock, Ted Anderson, Takeshi Amemiya, Mathias Cattaneo, Gary Chamberlain, Denis Chetverikov, Graham Elliott, Max Farrell, Eric Tchetgen, Bruce Hansen, James Hamilton, Jin Hahn, Han Hong, Guido Imbens, Zhipeng Liao, Tom MaCurdy, Anna Mikusheva, Whitney Newey, Alexei Onatsky, Joseph Romano, Shinichi Sakata, Andres Santos, Qi-man Shao, Mathew Shum, Chris Sims, Harrison Zhou, and participants of 10th Econometric World Congress in Shanghai in 2010, Caltech, CIREQ-Montreal, Harvard-MIT, UCL, UCLA, UC San-Diego, USC, UC-Davis, Princeton, Stanford, 2011 Infometrics Workshop, and Yale NSF 2012 conference on “High-Dimensional Statistics” for extremely helpful comments.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2011), “Bias-Corrected Matching Estimators for Average Treatment Effects”, *Journal of Business Econom. Statist.*, **29**, 1–11.

ANDREWS, D., CHENG, X. and GUGGENBERGER, P. (2011), “Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests” (Cowles Foundation Discussion Paper).

ANDREWS, D. and CHENG, X. (2011), “Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure” (Cowles Foundation Discussion Paper).

ANGRIST, J. D. and PISCHKE, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist’s Companion* (Princeton, New Jersey: Princeton University Press).

BACH, F. (2010), “Self-Concordant Analysis for Logistic Regression”, *Electronic Journal of Statistics*, **4**, 384–414.

BARANIUK, R., DAVENPORT, M., DEVORE, R. and WAKIN, M. (2008), “A Simple Proof of the Restricted Isometry Property for Random Matrices”, *Constructive Approximation*, **28**, 253–263.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012), “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain”, *Econometrica*, **80**, 2369–2429.

BELLONI, A. and CHERNOZHUKOV, V. (2011a), “ℓ_1-Penalized Quantile Regression in High-Dimensional Sparse Models”, *Ann. Statist.*, **39**, 82–130.

BELLONI, A. and CHERNOZHUKOV, V. (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.

BELLONI, A. and CHERNOZHUKOV, V. (2013), “Least Squares After Model Selection in High-dimensional Sparse Models”, *Bernoulli*, **19**, 521–547.

BELLONI, A., CHERNOZHUKOV, V., FERNANDEZ-VAL, I. and HANSEN, C. (2013), “Program Evaluation with High-Dimensional Data”.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2010), “LASSO Methods for Gaussian Instrumental Variables Models”, arXiv:[math.ST], http://arxiv.org/abs/1012.1297.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2011), “Inference for High-Dimensional Sparse Econometric Models”, *Advances in Economics and Econometrics. 10th World Congress of Econometric Society*.

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2013a), “Uniform Post Selection Inference for LAD Regression Models”, *arXiv preprint arXiv:1304.0282*.

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2010), “Square-Root-LASSO: Pivotal Recovery of Nonparametric Regression Functions via Conic Programming” (Duke and MIT Working Paper).

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011), “Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming”, *Biometrika*, **98**, 791–806.

BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2013b), “Honest Confidence Regions for Logistic Regression with a Large Number of Controls”, *arXiv preprint arXiv:1304.3969*.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009), “Simultaneous Analysis of Lasso and Dantzig Selector”, *Annals of Statistics*, **37**, 1705–1732.

CANDÈS, E. and TAO, T. (2007), “The Dantzig Selector: Statistical Estimation When p is much Larger than n”, *Ann. Statist.*, **35**, 2313–2351.

CATTANEO, M., JANSSON, M. and NEWEY, W. (2010), “Alternative Asymptotics and the Partially Linear Model with Many Regressors” (Working Paper, http://econ-www.mit.edu/files/6204).

CATTANEO, M. D. (2010), “Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability”, *Journal of Econometrics*, **155**, 138–154.

CHEN, X. (2007), “Large Sample Sieve Estimation of Semi-Nonparametric Models”, *Handbook of Econometrics*, **6**, 5559–5632.

CHEN, X. and POUZO, D. (2009), “Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals”, *Journal of Econometrics*, **152**, 46–60.

CHEN, X. and POUZO, D. (2012), “Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments”, *Econometrica*, **80**, 277–322.

DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009), *Self-Normalized Processes*, Probability and its Applications (New York). Limit theory and statistical applications (Berlin: Springer).

DONALD, S. G. and NEWEY, W. K. (1994), “Series Estimation of Semilinear Models”, *J. Multivariate Anal.*, **50**, 30–40.

DONOHUE III, J. J. and LEVITT, S. D. (2001), “The Impact of Legalized Abortion on Crime”, *Quarterly Journal of Economics*, **116**, 379–420.

DONOHUE III, J. J. and LEVITT, S. D. (2008), “Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz”, *Quarterly Journal of Economics*, **123**, 425–440.

DUFLO, E., GLENNERSTER, R. and KREMER, M. (2008), “Using Randomization in Development Econmics Research: A Toolkit”, in Schultz, T. P. and Strauss, J. A. (eds) *Handbook of Development Economics*, Vol. 4 (Elsevier: North-Holland).

FAN, J. and LI, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties”, *Journal of American Statistical Association*, **96**, 1348–1360.

FARRELL, M. (2013), “Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations”, (Working Paper, University of Michigan, August 2013).

FOOTE, C. L. and GOETZ, C. F. (2008), “The Impact of Legalized Abortion on Crime: Comment”, *Quarterly Journal of Economics*, **123**, 407–423.

FRANK, I. E. and FRIEDMAN, J. H. (1993), “A Statistical View of Some Chemometrics Regression Tools”, *Technometrics*, **35**, 109–135.

GAUTIER, E. and TSYBAKOV, A. (2011), “High-dimensional Instrumental Variables Rergession and Confidence Sets” arXiv:1105.2454v2 [math.ST].

HAHN, J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects”, *Econometrica*, **66**, 315–331.

HANSEN, B. E. (2005), “Challenges for Econometric Model Selection”, *Econometric Theory*, **21**, 60–68.

HÄRDLE, W., LIANG, H. and GAO, J. (2000), *Partially Linear Models*, Contributions to Statistics (Heidelberg: Physica-Verlag).

HECKMAN, J., LALONDE, R. and SMITH, J. (1999), “The Economics and Econometrics of Active Labor Market Programs”, *Handbook of Labor Economics*, **3**, 1865–2097.

HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998), “Matching as an Econometric Evaluation Estimator”, *Rev. Econom. Stud.*, **65**, 261–294.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003), “Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score”, *Econometrica*, 71(4), 1161–1189.

HUANG, J., HOROWITZ, J. L. and MA, S. (2008), “Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models”, *The Annals of Statistics*, **36**, 587—613.

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010), “Variable Selection in Nonparametric Additive Models”, *Ann. Statist.*, **38**, 2282–2313.

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

JING, B.-Y., SHAO, Q.-M. and WANG, Q. (2003), “Self-Normalized Cramér-type Large Deviations for Independent Random Variables”, *Ann. Probab.*, **31**, 2167–2215.

KERKYACHARIAN, G. and PICARD, D. (1992), “Density Estimation in Besov Spaces”, *Statist. Probab. Lett.*, **13**, 15–24.

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

KREMER, M. and GLENNERSTER, R. (2011), “Improving Health in Developing Countries: Evidence from Randomized Evaluations”, in Pauly, M. V., McGuire, T. G. and Barros, P. P. (eds) *Handbook of Health Economics*, Vol. 2 (North-Holland: Elsevier).

LEEB, H. and PÖTSCHER, B. M. (2008a), “Can one Estimate the Unconditional Distribution of Post-Model-Selection Estimators?”, *Econometric Theory*, **24**, 338–376.

LEEB, H. and PÖTSCHER, B. M. (2008b), “Recent Developments in Model Selection and Related Areas”, *Econometric Theory*, **24**, 319–322.

MACKINNON, J. G. and WHITE, H. (1985), “Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties”, *Journal of Econometrics*, **29**, 305–325.

MEINSHAUSEN, N. and YU, B. (2009), “Lasso-type Recovery of Sparse Representations for High-Dimensional Data”, *Annals of Statistics*, **37**, 2246–2270.

MIKUSHEVA, A. (2007), “Uniform Inference in Autoregressive Models”, *Econometrica*, **75**, 1411–1452.

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

NEYMAN, J. (1979), “$C(\alpha)$ Tests and their Use”, *Sankhya*, **41**, 1–21.

PÖTSCHER, B. M. (2009), “Confidence Sets based on Sparse Estimators are Necessarily Large”, *Sankhyā*, **71**, 1–18.

ROBINSON, P. M. (1988), “Root-N-Consistent Semiparametric Regression”, *Econometrica*, **56**, 931–954.

ROBINS, J. M. and ROTNITZKY, A. (1995), “Semiparametric Efficiency in Multivariate Regression Models with Missing Data”, *J. Amer. Statist. Assoc.*, **90**, 122–129.

ROMANO, J. P. (2004), "On Non-Parametric Testing, the Uniform Behaviour of the t-test, and Related Problems", *Scand. J. Statist.*, **31**, 567–584.

ROTHE, C. and FIRPO, S. (2013), "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions" (Discussion paper, NYU preprint).

RUDELSON, M. and ZHOU, S. (2011), "Reconstruction from Anisotropic Random Measurements", *ArXiv:1106.1151*.

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

VAN DE GEER, S. A. (2008), "High-Dimensional Generalized Linear Models and the Lasso", *Annals of Statistics*, **36**, 614–645.

VON BAHR, B. and ESSEEN, C.-G. (1965), "Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leqslant r \leqslant 2$", *Ann. Math. Statist*, **36**, 299–303.

ZHOU, S. (2009), "Restricted Eigenvalue Conditions on Subgaussian Matrices", *ArXiv:0904.4723v2*.