INFERENCE ON TREATMENT EFFECTS AFTER SELECTION AMONGST HIGH-DIMENSIONAL CONTROLS

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

Abstract. We propose robust methods for inference on the effect of a treatment variable on a scalar outcome in the presence of very many controls. Our setting is a partially linear model with possibly non-Gaussian and heteroscedastic disturbances. Our analysis allows the number of controls to be much 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 conditioning on a relatively small number of controls whose identities are unknown. The latter condition makes it possible to estimate the treatment effect by selecting approximately the right set of controls. We develop a novel estimation and uniformly valid inference method for the treatment effect in this setting, called the "post-double-selection" method. Our results apply to Lasso-type methods used for covariate selection as well as to any other model selection method that is able to find a sparse model with good approximation properties.

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 long-standing problem of uniform inference after model selection for a large, interesting class of models. We illustrate the use of the developed methods with numerical simulations and an application to 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

1. Introduction

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

Date: First version: May 2010. This version is of April 12, 2012. This is a revision of an Dec 2011 ArXiv/CEMMAP paper entitled "Estimation of Treatment Effects with High-Dimensional Controls".

We thank Stéphane Bonhomme, Mathias Cattaneo, Gary Chamberlain, Denis Chetverikov, Eric Gretchen, Bruce Hansen, Guido Imbens, Anna Mikusheva, Whitney Newey, Andres Santos, and participants of 10th Econometric World Congress in Shanghai, Harvard-MIT, UC San-Diego, Princeton, and Infometrics Workshop for helpful comments.

estimating the causal effect of the minimum wage or some other government policy on employment. Since economic policies and many other economic variables are not randomly assigned, economists rely on a variety of quasi-experimental approaches based on observational data when trying to estimate such effects. One important method is based on the assumption that the variable of interest can be taken as randomly assigned once a sufficient set of other factors has been controlled for. Economists, for example, might argue that changes in state-level minimum wages can be taken as randomly assigned relative to unobservable factors that could affect changes in state-level employment once aggregate macroeconomic activity, state-level economic activity, and state-level demographics have been controlled for; see Card and Krueger (1997), Heckman, LaLonde, and Smith (1999), Imbens (2004), among other references.

A problem empirical researchers face when relying on an identification strategy for estimating a structural effect that relies on a conditional on observables argument is knowing which 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 attempting to select a sensible set of controls from a potentially vast set of control variables including raw regressors available in the data as well as interactions and other transformations of these regressors. A typical economic study will rely on 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 paper, 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 $\mathrm{E}[\zeta_i \mid d_i, z_i] = 0$.[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 controls $x_i = P(z_i)$, 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 on the data. We impose such

[1] We note that d_i does not need to be binary.

structure by assuming that there is a relatively small set consisting of $s < n$ variables whose identities are *a priori* unknown among the p potential conditioning variables that provides a good enough approximation to $g(z_i)$ that the exogeneity of d_i may be taken as given once these variables have been controlled for linearly. 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 allows for the realistic scenario in which the researcher is unsure about exactly which variables or transformations are important for approximating $g(z_i)$ and so must search among a broad set of controls.

The assumed sparsity includes as special cases the most common approaches to parametric and nonparametric regression analysis. Sparsity justifies the use of fewer variables than there are observations in the sample. When the initial number of variables is high, the assumption justifies the use of variable selection methods to reduce the number of variables to a manageable size. In many economic applications, formal and informal strategies are often used to select such smaller sets of potential control variables. Most of these standard variable selection strategies are highly non-robust, leading to a very poor inference.[2] In an effort to demonstrate robustness, researchers often employ *ad hoc* sensitivity analyses which examine the robustness of inferential conclusions to variations in the set of controls. Such sensitivity analyses are useful but lack rigorous justification. As a complement to these *ad hoc* approaches, we propose a formal, rigorous approach to inference allowing for selection of controls. Our proposal uses modern variable selection methods in a novel manner which results in robust and valid inference.

The main contributions of this paper are providing a robust estimation and inference method within a partially linear model with potentially very high-dimensional controls and developing the supporting theory. The method relies on the use of Lasso-type or other sparsity-inducing procedures for variable selection. Our approach sharply 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 robustness 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

[2] An example of inference going wrong is given in Figure 1 (left panel), presented in the next section, where the standard post-model selection estimator has a bimodal distribution which sharply deviates from the standard normal distribution. More examples are given in Section 6 where we document the poor inferential performance of the standard post-model selection methods.

the equation of interest, ideally helping keep the residual variance small as well as intuitively 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.[3]

We illustrate the theoretical results through an examination of the effect of abortion on crime rates following Donohue III and Levitt (2001), Foote and Goetz (2008), and Donohue III and Levitt (2008). In the original data of Donohue III and Levitt (2001), we find that the formal variable selection procedure applied to a set of variables that allows for parsimonious but flexible trends produces a qualitatively different result than that obtained through the *ad hoc* set of sensitivity results presented in the original paper. Interestingly, we come to a similar conclusion as in Foote and Goetz (2008) that the estimated abortion effects becomes quite imprecise once a small set of variables including state-level average abortion rates interacted with a linear trend are included in the model. We also consider a model related to Donohue III and Levitt (2008) and once again find that the variable selection procedures choose a small number of controls that allow for simple nonlinear trends in the data interacted with fixed, state-level variables and that the estimated abortion effect is rendered imprecise once these variables are included in the model of Donohue III and Levitt (2008). The selection of these variables and the resulting imprecision of the estimated treatment effect suggest that one cannot determine precisely whether the effect attributed to abortion found when these parsimonious trend terms are omitted from the model is due to changes in the abortion rate or some other persistent state-level factor that is related to relevant changes in the abortion rate and current changes in the crime rate.[4] Finding that a simple-to-implement, formal 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

[3]In a companion paper that presents an overview of results for ℓ_1-penalized estimators, Belloni, Chernozhukov, and Hansen (2011a), we provide similar results in the idealized Gaussian homoscedastic framework.

[4]Note that all models are estimated including fixed effects or more general sets of controls to eliminate any state-specific factors that might be related to both the relevant level of the abortion rate and the level of the crime rate.

estimating treatment effects in settings where treatment is taken as exogenous conditional on observables.

Relationship to literature. We contribute to several existing literatures. First, we contribute to the literature on series estimation of partially linear models (Donald and Newey (1994), Härdle, Liang, and Gao (2000), Robinson (1988), and others). We differ from most of the existing literature by considering many series terms, $p \gg n$, that may be used in the construction of the regression fits. Considering a broad set of terms allows for more refined approximations of regression functions relative to the usual approach that only allows for a few low-order terms. Simultaneously, the number of parameters is kept relatively low through the use of variable selection methods. See, for example, Belloni, Chernozhukov, and Hansen (2011a) for a wage function example and Section 5 for theoretical examples. Moreover, we allow for data-dependent selection of the appropriate series terms. We focus on Lasso as our principal device for performing this selection but note that any other method, such as selection using the traditional generalized cross-validation criteria, is likely to work as long as the method guarantees sufficient sparsity in its solution. It should be noted that the previous literature on inference in the partially linear model generally takes the number of series terms given without allowing for their data-driven selection. Recent work by Cattaneo, Jansson, and Newey (2010) allows for $p = Cn$ (with $C < 1$) series terms in construction of the series regression, a device which produces refined standard errors even in cases with $p \ll n$. This work is complementary to our work which focuses on reducing the number of terms $p \gg n$ down to $\widehat{s} \ll n$. After model selection, one may apply conventional standard errors or Cattaneo, Jansson, and Newey (2010) standard errors.[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. 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 connection of our approach to the propensity score brings about interesting connections to the treatment effects literature. Hahn (1998), Heckman, Ichimura, and

[5]In practice, if the selected number of terms $\widehat{s}$ is substantial, we recommend using Cattaneo, Jansson, and Newey (2010)'s standard errors after applying our model selection procedure.

Todd (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, Imbens, and Ridder (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,[6] but there has been little work on inference after imperfect model selection. Perfect model selection relies on unrealistic assumptions, and model selection mistakes can have serious consequences for inference as has been shown in Pötscher (2009), Leeb and Pötscher (2008), and others. In work on instrument selection for estimation of a linear instrumental variables model, Belloni, Chen, Chernozhukov, and Hansen (2010) have shown that model selection mistakes do not prevent valid inference about low-dimensional structural parameters due to the inherent adaptivity of the problem: Omission of a relevant instrument does not affect consistency of an IV estimator as long as there is another relevant instrument. The partially linear regression model (1.1) does not have the same adaptivity structure, and model selection based on the outcome regression alone produces non-robust confidence intervals. Our procedure creates the necessary adaptivity by performing two separate model selection steps, making it possible to perform robust/uniform inference after model selection. The uniformity holds over large, interesting classes of models. In that regard, our contribution is in the spirit of recent contributions 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, Cheng, and Guggenberger (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-penalization 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 Tibshirani (1996), in part because they are computationally efficient. Many ℓ_1-penalized methods have been shown to have good

[6]For reviews focused on econometric applications, see, e.g., Hansen (2005) and Belloni, Chernozhukov, and Hansen (2010).

estimation properties even when perfect variable selection is not feasible; see, e.g., Candes and Tao (2007), Meinshausen and Yu (2009), Bickel, Ritov, and Tsybakov (2009), Belloni and Chernozhukov (2011) and the references therein. Such methods have also been shown to extend suitably to nonparametric and non-Gaussian cases as in Bickel, Ritov, and Tsybakov (2009) and Belloni, Chen, Chernozhukov, and Hansen (2010). These methods also 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.

Paper Organization. In Section 2, we formally present the modeling environment including the key sparsity condition and develop our advocated estimation and inference method. We establish the consistency and asymptotic normality of our estimator of α_0 uniformly over large classes of models in Section 3. In Section 4, we present a generalization of the basic procedure to allow for model selection methods other than Lasso. In Section 4, we present a series of theoretical examples in which we provide primitive condition that imply the higher-level conditions of Section 3. We present a series of numerical examples that verify our theoretical results numerically in Section 6, and we apply our method to the abortion and crime example of Donohue III and Levitt (2001) in Section 7. In appendices, we provide the proofs.

Notation. In what follows, we work with triangular array data $\{(\omega_{i,n}, i = 1, ..., n), n = 1, 2, 3, ...\}$ defined on probability space $(\Omega, \mathcal{A}, \mathrm{P}_n)$, where $\mathrm{P} = \mathrm{P}_n$ can change with n. 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 $\{\omega_{i,n}, i = 1, ..., n\}$ are implicitly indexed by P_n and thus by n. We omit the dependence on these objects from the notation in what follows for notational simplicity. We use array asymptotics to 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 of data-generating processes.

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, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use $\|\cdot\|_\infty$ to denote the maximal element of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant

$c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E, we say that E wp $\to 1$ when E occurs with probability approaching one as n grows. Given a p-vector b, we denote $\text{support}(b) = \{j \in \{1,...,p\} : b_j \neq 0\}$. Throughout the paper, 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 will mean constants that are given and that do not depend on $\mathrm{P} = \mathrm{P}_n$.

2. Inference on Treatment and Structural Effects Conditional on Observables

2.1. **Framework.**

In this paper we consider the following partially linear model

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

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

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

The confounding factors z_i affect the policy variable via the function $m(z_i)$ and the outcome variable via 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 a dimension $p = p_n$ which can be large relative to the sample size. Specifically, our results only 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) or Chen (2007).

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} sufficiently 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 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.** We propose the following method for estimating and performing inference about α_0. 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.[7] The construction of our advocated procedure reflects our effort to offer a method that has attractive robustness/uniformity properties for inference. The estimator is $\sqrt{n}$-consistent and asymptotically normal under mild conditions and provides confidence intervals that are robust to various perturbations of the data-generating process that preserve approximate sparsity.

[7]To the best of our knowledge, our result is the first of its kind in this setting. This result extends our previous results on inference under imperfect model selection in the instrumental variables model given in Belloni, Chernozhukov, and Hansen (2010). The problem is more difficult in the present paper due to lack of adaptivity in estimation which we overcome by introducing additional model selection steps.

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. The chief method we discuss is the Lasso method described in more detail below. 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. 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 efficiency bound[8] than the single selection method.

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

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

The set $\widehat{I}$ may contain variables that were not selected in the variable selection steps with indices in $\widehat{I}_3$ that the analyst thinks are important for ensuring robustness. We call $\widehat{I}_3$ the amelioration set. Thus, $\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2 \cup \widehat{I}_3$; let $\widehat{s} = |\widehat{I}|$ and $\widehat{s}_j = |\widehat{I}_j|$ for $j = 1, 2, 3$.

We define feasible Lasso estimators below and note that other selection methods could be used as well. When a feasible Lasso is used, we refer to the post-double-selection estimator as the post-double-Lasso estimator.

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

$$([\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0, 1) \tag{2.8}$$

[8] Semi-parametric efficiency is attained in the homoscedastic case.

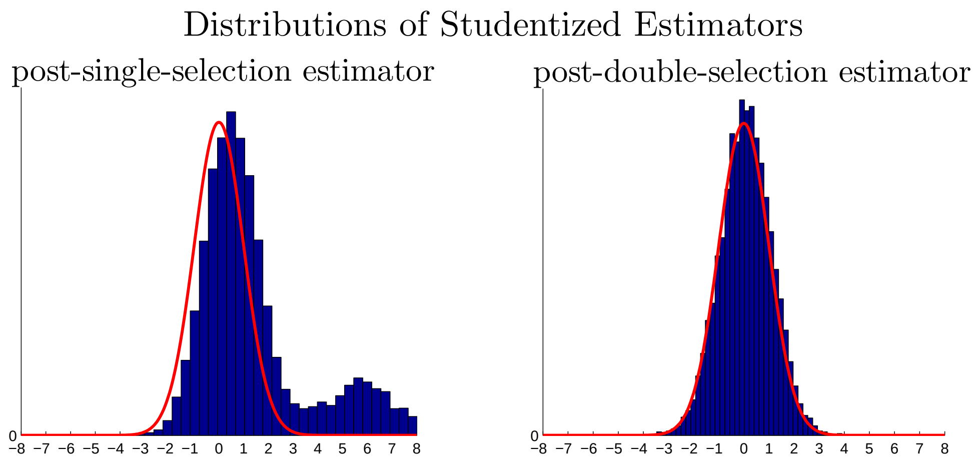

FIGURE 1. The finite-sample distributions (densities) of the standard post-single selection estimator (left panel) and of our proposed post-double selection estimator (right panel). The distributions are given for centered and studentized quantities. The results are based on 10000 replications of Design 1 described in Section 6, with R^2's in equation (2.6) and (2.7) set to 0.5.

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. All of these results imply uniform validity of confidence regions over large, interesting classes of models. Figure 2.2 (right panel) illustrates the result (2.8) by showing that the finite-sample distribution of our post-double-selection estimator is very close to the normal distribution. In contrast, Figure 2.2 (left panel) illustrates the classical 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. Tibshirani (1996) proposes the Lasso estimator which is defined as a solution to

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

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. The kinked nature of the penalty function induces the solution $\widehat{\beta}$ to have many zeroes and thus may be used for model selection. The selected model $\widehat{T} = \text{support}(\widehat{\beta})$ is often used for further refitting by least squares, leading to the so called post-Lasso or Gauss-Lasso estimator, see, e.g., Belloni and Chernozhukov (2011). The Lasso estimator/selector is computationally attractive because it minimizes a convex function. In the homoskedastic Gaussian case, a basic choice for penalty level suggested by Bickel, Ritov, and Tsybakov (2009) is

$$\lambda = 2 \cdot c\sigma\sqrt{2n\log(2p/\gamma)}, \tag{2.10}$$

where $c > 1$, $1-\gamma$ is a confidence level that needs to be set close to 1, and σ is the standard deviation of the noise. The formal motivation for this penalty is that it leads to near-optimal rates of convergence of the estimator under approximate sparsity. The good behavior of the estimator of β_0 in turn implies good approximation properties of the selected model $\widehat{T}$, as noted in Belloni and Chernozhukov (2011). Unfortunately, even in the homoskedastic case the penalty level specified above is not feasible since it depends on the unknown σ.

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

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

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

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

where $c > 1$ and $1-\gamma$ is a confidence level.[9] The l_j's are the 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. We refer to the resulting feasible Lasso method as the *Iterated Lasso*. The estimator $\widehat{\beta}$ has statistical performance that is similar to that of the (infeasible) Lasso described above in Gaussian cases and delivers similar performance in non-Gaussian, heteroscedastic cases; see Belloni, Chen, Chernozhukov, and Hansen (2010). In this paper, we only 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.

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

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

$$\min_{\beta\in\mathbb{R}^p} \sqrt{\mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2]} + \frac{\lambda}{n}\|\widehat{\Psi}\beta\|_1, \tag{2.13}$$

with the penalty level

$$\lambda = c \cdot \sqrt{n}\Phi^{-1}(1-\gamma/2p), \tag{2.14}$$

where $c > 1$, $\gamma \in (0,1)$ is a confidence level, and $\widehat{\Psi} = \text{diag}(\widehat{l}_1, \dots, \widehat{l}_p)$ is a diagonal matrix of penalty loadings. The main attractive feature of (2.13) is that in the homoscedastic case we can set $\widehat{l}_j = \{\mathbb{E}_n[\tilde{x}_{ij}^2]\}^{1/2}$, which depends only on observed data and is thus very convenient.

In the heteroscedastic case, we would like to choose $\widehat{l}_j$ so that

$$l_j + o_P(1) \leqslant \widehat{l}_j \lesssim_P l_j, \text{ where } l_j = \{\mathbb{E}_n[\tilde{x}_{ij}^2\epsilon_i^2]/\mathbb{E}_n[\epsilon_i^2]\}^{1/2}, \text{ uniformly in } j = 1, ..., p. \tag{2.15}$$

As a simple bound, we could use $\widehat{l}_j = 2\{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/4}$ since

$$\{\mathbb{E}_n[\tilde{x}_{ij}^2\epsilon_i^2]/\mathbb{E}_n[\epsilon_i^2]\}^{1/2} \leqslant \{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/4}\{\mathbb{E}_n[\epsilon_i^4]\}^{1/4}/\{\mathbb{E}_n[\epsilon_i^2]\}^{1/2}.$$

This bound gives $l_j + o_P(1) \leqslant \widehat{l}_j$ if $\{\mathbb{E}_n[\epsilon_i^4]\}^{1/4}/\{\mathbb{E}_n[\epsilon_i^2]\}^{1/2} \leqslant 2 + o_P(1)$, which covers a wide class of marginal distributions for error ϵ_i. For example, all t-distributions with degrees of freedom greater than five satisfy this condition. As in the previous case, we can also iteratively re-estimate the penalty loadings using estimates of the ϵ_i's to approximate the ideal penalty loadings:

$$\widehat{l}_j = l_j + o_P(1), \text{ uniformly in } j = 1, ..., p. \tag{2.16}$$

The resulting Square-root Lasso and post-Square-root Lasso estimators based on these penalty loadings achieve near optimal rates of convergence even in drastically non-Gaussian, heteroscedastic cases. This good performance implies good approximation properties for the selected model $\widehat{T}$.

In what follows, we shall use the term *feasible Lasso* to refer to either the Iterated Lasso estimator $\widehat{\beta}$ solving (2.11)-(2.12) or the Square-root Lasso estimator $\widehat{\beta}$ solving (2.13)-(2.15) with $c > 1$ and $1 - \gamma$ set such that

$$\gamma = o(1) \text{ and } \log(1/\gamma) \lesssim \log(p \vee n). \tag{2.17}$$

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 as well as other more technical assumptions.

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

Condition ASTE (P). *(i)* $\{(y_i, d_i, z_i), i = 1, ..., n\}$ *are i.n.i.d. vectors on* $(\Omega, \mathcal{F}, \mathrm{P})$ *that obey the model (2.2)-(2.3), and the vector* $x_i = P(z_i)$ *is a 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 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*

$$m(z_i) = x_i'\beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{mi}^2]\}^{1/2} \leqslant C\sqrt{s/n}, \tag{3.18}$$

$$g(z_i) = x_i'\beta_{g0} + r_{gi}, \qquad \|\beta_{g0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{gi}^2]\}^{1/2} \leqslant C\sqrt{s/n}. \tag{3.19}$$

(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 the growth 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 $\widehat{I}_3$ should 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 $\widehat{s} \lesssim_P s$, and we also require that $s^2 \log^2(p \vee n)/n \to 0$. This condition can be relaxed using the sample-splitting method of Fan, Guo, and Hao (2011), which is done in the 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 behavior of the Gram matrix $\mathbb{E}_n[x_i x_i']$. Whenever $p > n$, the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ does not have full rank and in principle is not well-behaved. However, we only need good behavior of smaller submatrices. Define the minimal and maximal m-sparse eigenvalue of a semi-definite matrix M as

$$\phi_{\min}(m)[M] := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2} \text{ and } \phi_{\max}(m)[M] := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}. \tag{3.20}$$

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). *There is an absolute sequence of constants $\ell_n \to \infty$ such that the maximal and minimal $\ell_n s$-sparse eigenvalues are bounded from below 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, $i = 1, \ldots, n$, are i.i.d. zero-mean sub-Gaussian random vectors that have population Gram matrix $\mathrm{E}[x_i x_i']$ 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, $i = 1, \ldots, n$, are i.i.d. bounded zero-mean random vectors with $\|x_i\|_\infty \leqslant K_n$ a.s. that have population Gram matrix $\mathrm{E}[x_i x_i']$ with minimal and maximal $s \log n$-sparse eigenvalues bounded from above and away from zero by absolute constants, where $K_n^2 s(\log^3 n)\{\log(p \vee n)\}/n \leqslant \delta_n \to 0$.

The claim (a) holds by Theorem 3.2 in Rudelson and Zhou (2011) (see also Zhou (2009) and Baraniuk, Davenport, DeVore, and Wakin (2008)) and claim (b) holds by Lemma 1 in Belloni and Chernozhukov (2011) or by Theorem 1.8 Rudelson and Zhou (2011). Recall that a standard assumption in econometric research is to assume that the population Gram matrix $\mathrm{E}[x_i x_i']$ has eigenvalues bounded from above and away from zero, see e.g. Newey (1997). The conditions above allow for this and more general behavior, requiring only that the $s \log n$ sparse eigenvalues of the population Gram matrix $\mathrm{E}[x_i x_i']$ are bounded from below and from above. □

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

Condition SM (P). *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) $\bar{\mathrm{E}}[|d_i|^q] \leqslant C, \quad c \leqslant \mathrm{E}[\zeta_i^2 \mid x_i, v_i] \leqslant C$ *and* $c \leqslant \mathrm{E}[v_i^2 \mid x_i] \leqslant C$ *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, which are rather mild, 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, Shao, and Wang (2003) to bound some important error components.

3.2. **The Main Result.** The following is the main result of this paper. It shows that the post-double selection estimator is root-n consistent and 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* $\check{\alpha}$ *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* $\widehat{v}_i := d_i - x_i'\widehat{\beta}$*,* $i = 1,\ldots,n$ *where* $\widehat{\beta} \in \arg\min_\beta\{\mathbb{E}_n[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}\}$.

An immediate consequence of this result is the following corollary.

Corollary 1 (**Uniformly Valid Confidence Intervals**)**.** *(i) 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*. Let* $c(1-\gamma) = \Phi^{-1}(1-\gamma/2)$*. Then as* $n \to \infty$*, uniformly in* $\mathrm{P} \in \mathbf{P}_n$

$$\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm c(1-\gamma)\widehat{\sigma}_n/\sqrt{n}]\right) \to 1-\gamma.$$

(ii) 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* $n \geqslant n_0$ *for some* n_0*. Then as* $n \to \infty$*, uniformly in* $\mathrm{P} \in \mathbf{P}$

$$\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm c(1-\gamma)\widehat{\sigma}_n/\sqrt{n}]\right) \to 1-\gamma.$$

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. These regions appear to be substantial, as we demonstrate via a sequence of theoretical and numerical examples in Section 5 and 6. In contrast, the standard post-selection method based on (2.4) generates non-robust confidence intervals.

Comment 3.3. Our approach to uniformity analysis 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 most 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. Hence our approach seems to be quite different in spirit to the generic uniformity analysis suggested by Andrews, Cheng, and Guggenberger (2011). □

Comment 3.4. Uniformity holds over a large class of approximately sparse models, which cover conventional models used in series estimation of partially linear models as shown in Section 5. Of course, for every interesting class of models and any inference method, one could find an even bigger class of models, where the uniformity does not apply, which is also true in our case. In particular our models do not cover models with many small coefficients (in the series case, this corresponds to small deviations from smoothness towards highly non-smooth functions, namely towards paths generated as realization of an approximate white noise process). This motivates further research work on inference procedures that have robustness properties to deviations from the given class of models that are deemed important. In the simulations in Section 6, we consider incorporating the ridge fit along the other controls to be selected over using lasso to build extra robustness against such deviations away from approximately sparse models. □

3.3. **Auxiliary Results on Model Selection via Lasso and Post-Lasso.**

The post-double-selection estimator applies the least squares estimator to the union of 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 section is to describe these properties. The proof of Theorem 1 relies on these properties.

Note that either of the regression models (2.6)-(2.7) obey the following conditions.

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

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

Let $\widehat{T}$ denote the model selected by the feasible Lasso estimator $\widehat{\beta}$. Formally, set

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

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

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \ \mathbb{E}_n[(\tilde{y}_i - \tilde{x}_i'\beta)^2] \quad : \quad \beta_j = 0 \text{ for each } j \notin \widehat{T}. \tag{3.21}$$

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

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

Condition RF. *The following conditions hold:*

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

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)**.** *Suppose that conditions ASM and RF hold, and that Condition SE holds for* $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$. *Consider the choices for penalty level and loadings described in Section 3.3 for a feasible Lasso estimator.*

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

$$\widehat{s} = |\widehat{T}| \lesssim s \quad wp \quad 1 - o(1)$$

and

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

(ii) The Post-Lasso estimator obeys

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

and

$$\|\widetilde{\beta} - \beta_0\| \lesssim_P \sqrt{\mathbb{E}_n[\{\tilde{x}_i'\widetilde{\beta} - \tilde{x}_i'\beta_0\}^2]} \lesssim_P \sigma\sqrt{\frac{s\log(p\vee n)}{n}}. \tag{3.22}$$

Lemma 1 was derived in Belloni, Chen, Chernozhukov, and Hansen (2010) for Iterated Lasso and by Belloni, Chernozhukov, and Wang (2010) for Square-root Lasso. These analyses built on the rate analysis of infeasible Lasso by Bickel, Ritov, and Tsybakov (2009) and on sparsity analysis and rate analysis of Post-Lasso by Belloni and Chernozhukov (2011). Lemma 1 shows that feasible Lasso methods select a model $\widehat{T}$ that provides a high-quality approximation to the regression function $f(\tilde{z}_i)$, i.e. they find a sparse model that can approximate the function at the "near-oracle" rate $\sqrt{s/n}\sqrt{\log(p\vee n)}$. If we knew the "best" approximating model

$T = \text{support}(\beta_0)$, we could achieve the "oracle" rate of $\sqrt{s/n}$. Note that Lasso methods generally will not recover T perfectly. Moreover, no method can recover T perfectly in general, except under the restrictive condition that all non-zero coefficients in β_0 are bounded away from zero by a factor that exceeds estimation error. We do not require this condition to hold in our results. All that we need is that the selected model $\widehat{T}$ can approximate the regression function well and that the size of the selected model, $\widehat{s} = |\widehat{T}|$, is of the same stochastic order as $s = |T|$. This condition holds in many cases in which some non-zero coefficients are close to zero.

The lemma above also shows that feasible Post-Lasso achieves the same near-oracle rate as feasible Lasso. The coincidence in rates occurs despite the fact that feasible Lasso will in general fail to correctly select the best-approximating model T as a subset of the variables selected; that is, $T \not\subseteq \widehat{T}$. The intuition for this result is that any components of T that feasible Lasso misses are unlikely to be important.

4. Generalization: Inference after Double Selection by a Generic Selection Method

The conditions provided so far are simply a set 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). *A model selector provides a possibly data-dependent set* $\widehat{I} \subset \{1, ..., p\}$ *of covariate names such that, with probability* $1 - \Delta_n$, $|\widehat{I}| \leqslant Cs$ *and*

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

Condition HLMS requires that with high probability the selected model $\widehat{I}$ is sparse and generates a good approximation for the functions g and m. In practice the set $\widehat{I}$ can be selected by a variety of sparse methods. Examples of such methods include the Dantzig selector (Candes and Tao, 2007), feasible Dantzig selector (Gautier and Tsybakov, 2011), Bridge estimator (Huang, Horowitz, and Ma, 2008), SCAD penalized least squares (Fan and Li, 2001) and thresholded Lasso (Belloni and Chernozhukov, 2011), to name a few. We emphasize that, similarly to the previous arguments, we allow for imperfect model selection.

The following result establishes the inferential properties of a generic model selection device $\widehat{I}$ that obeys Condition HLSM.

Theorem 2 (Estimation and Inference on Treatment Effects under High-Level Model Selection)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes that satisfies conditions ASTE (P), SM (P), and SE (P) for $\mathrm{P} = \mathrm{P}_n$ for each n. Suppose that the selected model $\widehat{I}$ satisfies condition HLSM for each n. Then the estimator $\check{\alpha}$ based on $\widehat{I}$ obeys*

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

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

An immediate consequence of this result is the following corollary.

Corollary 2 (**Uniformly Valid Confidence Intervals**)**.** *(i) Let $\mathbf{P}_n$ be the collection of all data-generating processes P for which conditions ASTE(P), SM (P), SE (P), and HLSM (P) hold for given n. Let $c(1-\gamma) = \Phi^{-1}(1-\gamma/2)$. Then as $n \to \infty$, uniformly in $\mathrm{P} \in \mathbf{P}_n$*

$$\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm c(1-\gamma)\widehat{\sigma}_n/\sqrt{n}]\right) \to 1-\gamma.$$

(ii) 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 $n \geqslant n_0$ for some n_0. Then as $n \to \infty$, uniformly in $\mathrm{P} \in \mathbf{P}$

$$\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm c(1-\gamma)\widehat{\sigma}_n/\sqrt{n}]\right) \to 1-\gamma.$$

5. Theoretical Examples

The purpose of this section is to give a sequence of examples – progressing from simple to somewhat involved – that highlight the range of the applicability and robustness of the proposed method. In these examples, we specify primitive conditions which cover a broad range of applications including nonparametric models and high-dimensional parametric models. We emphasize that our main regularity conditions cover even more general models which combine various features of these examples such as models with both nonparametric and high-dimensional parametric components.

In all examples, the model is

$$\begin{aligned} y_i &= d_i\alpha_0 + g(z_i) + \zeta_i, & \mathrm{E}[\zeta_i \mid z_i, v_i] = 0,\\ d_i &= m(z_i) + v_i, & \mathrm{E}[v_i \mid z_i] = 0, \end{aligned} \tag{5.23}$$

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.

We start out with a "trivial" example, in which the dimension p of the regressors is fixed. In practical terms this example approximates cases with p small compared to n. Though simple, this example is important as standard post-single-selection methods fail to provide confidence intervals that are valid uniformly in the underlying data-generating process in this context; see Leeb and Pötscher (2008). 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) as i.i.d. vectors for $i = 1, ..., n$ obeying the model (5.23) with

$$\begin{array}{rl} g(z_i) = & \sum_{j=1}^{p} \beta_{g0j} z_{ij}, \\ m(z_i) = & \sum_{j=1}^{p} \beta_{m0j} z_{ij}. \end{array} \tag{5.24}$$

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

$$\begin{array}{l} \mathrm{E}[\|x_i\|^{q_x}] \leqslant B, \quad \|\alpha_0\| + \|\beta_{g0}\| + \|\beta_{m0}\| \leqslant B, \quad b \leqslant \lambda_{\min}(\mathrm{E}[x_i x_i']), \\ b \leqslant \mathrm{E}[\zeta_i^2 \mid x_i, v_i], \quad \mathrm{E}[|\zeta_i^q| \mid x_i, v_i] \leqslant B, \quad b \leqslant \mathrm{E}[v_i^2 \mid x_i], \quad \mathrm{E}[|v_i^q| \mid x_i] \leqslant B. \end{array} \tag{5.25}$$

Let $\mathbf{P}$ be the collection of all regression models P that obey the conditions set forth above for all n for the given constants (p, b, B, q_x, q). Then, as established in Appendix F, 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 $\mathrm{P} \in \mathbf{P}$. □

The next three examples are more substantial and include infinite-dimensional models which we approximate with linear functional forms with potentially very many regressors. The key to estimation in these models is a smoothness condition which requires regression coefficients to decay at some rates. In series and sieve 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

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

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). 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, Chernozhukov, and Hansen (2011a).

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 a.s. 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 coefficient 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

$$\beta_{h0} := \arg \min_{\|\beta\|_0 \leqslant s} \mathrm{E}[(h(z_i) - x_i'\beta)^2], \tag{5.26}$$

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) as i.i.d. vectors for $i = 1, ..., n$ obeying the model (5.23) with

$$\begin{aligned} g(z_i) &= \textstyle\sum_{j=1}^{\infty} \theta_{gj} z_{ij}, \\ m(z_i) &= \textstyle\sum_{j=1}^{\infty} \theta_{mj} z_{ij}. \end{aligned} \tag{5.27}$$

Assume that the infinite dimensional vector $w_i = (z_i, \zeta_i, v_i)$ is jointly Gaussian with zero mean and minimal and maximal eigenvalues of the covariance matrix (operator) $\mathrm{E}[w_i w_i']$ 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 the smoothness condition on the coefficients. Let $a > 1$ and $0 < A < \infty$ be some absolute constants. We require that the coefficients of the expansions in (5.27) are a-smooth with constant A after p-rearrangement, namely

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

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

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

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

Let $\mathbf{P}_n$ be the collection of all dgp P that obey the conditions set forth in this example for a given n and for the given constants $(\underline{\kappa}, \overline{\kappa}, a, A, B, \chi)$ and sequences p_n and $\bar{\delta}_n$. Then, as established in Appendix F, 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 $(\kappa', \kappa'', c, C)$ and sequences Δ_n and δ_n in those conditions depending only on $(\underline{\kappa}, \bar{\kappa}, a, A, B, \chi)$. 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 for any $\mathrm{P} \in \mathbf{P}_n$. In particular, these conclusions 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) as i.i.d. vectors for $i = 1, ..., n$ obeying the model:

$$\begin{aligned} g(z_i) = & \sum_{j=1}^{\infty} \theta_{gj} P_j(z_i), \\ m(z_i) = & \sum_{j=1}^{\infty} \theta_{mj} P_j(z_i), \end{aligned} \tag{5.28}$$

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 various compactly supported wavelets or canonical trigonometric bases.

Let $a > 1$ and $0 < A < \infty$ be some absolute constants. We require that the coefficients of the expansions in (5.28) are a-smooth with constant A after p-rearrangement, namely

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

This condition is directly connected to smoothness of the underlying function. 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).

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

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

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

$$\|\alpha_0\| \leqslant B, \ b \leqslant \mathrm{E}[\zeta_i^2 \mid x_i, v_i], \ \ \mathrm{E}[|\zeta_i^q| \mid x_i, v_i] \leqslant B, \ \ b \leqslant \mathrm{E}[v_i^2 \mid x_i], \ \ \mathrm{E}[|v_i^q| \mid x_i] \leqslant B. \tag{5.29}$$

Let $\mathbf{P}_n$ be the collection of all regression models P that obey the conditions set forth above for a given n. Then, as established in Appendix F, 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 the constants $(\underline{f}, \bar{f}, a, A, b, B, q)$. 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 for any $\mathrm{P} \in \mathbf{P}_n$. In particular, as a special case, the same conclusion applies uniformly in $\mathrm{P} \in \mathbf{P} = \cap_{n \geqslant n_0} \mathbf{P}_n$. □

The following example generalizes the previous example by allowing for sieve, as opposed to series, approximations.

Example 4. (Sieve 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) as i.i.d. vectors for $i = 1, ..., n$ obeying the model (5.23):

$$\begin{aligned} g(z_i) &= \textstyle\sum_{j=1}^{p} \theta_{gj} P_{j,p}(z_i) + \rho_{g,p}(z_i), \\ m(z_i) &= \textstyle\sum_{j=1}^{p} \theta_{mj} P_{j,p}(z_i) + \rho_{m,p}(z_i), \end{aligned} \tag{5.30}$$

where $\rho_{g,p}(z_i)$ and $\rho_{m,p}(z_i)$ are approximation errors. We assume that z_i has support $[0,1]^d$ with density bounded from below by constant $\underline{f} > 0$ and above by constant $\bar{f}$, and $P_p(z_i) = (P_{j,p}(z_i), j = 1, ..., p)$ is an orthonormal sub-basis on $L^2[0,1]^d$ with bounded elements, i.e. $\max_{z \in [0,1]^d} |P_{j,p}(z)| \leqslant B$ for all $j = 1, 2, ...$ Here all constants are taken to be absolute.

The main assumption that guarantees approximate sparsity is the smoothness condition on the coefficients. Let $a > 1$ and $0 < A < \infty$ be some absolute constants. We require that the coefficients of the expansions in (5.30) are a-smooth with constant A after p-rearrangement, namely

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

We also impose that the approximation errors obey the conditions:

$$\begin{array}{ll} \sqrt{\mathrm{E}[\rho^2_{g,p}(z_i)]} \leqslant Ap^{-a+1/2}, & \sup_{z\in[0,1]^d} |\rho_{g,p}(z)| \leqslant Ap^{-a+1}, \\ \sqrt{\mathrm{E}[\rho^2_{m,p}(z_i)]} \leqslant Ap^{-a+1/2}, & \sup_{z\in[0,1]^d} |\rho_{m,p}(z)| \leqslant Ap^{-a+1}. \end{array} \tag{5.31}$$

These assumptions are in line with standard assumptions on sieve approximations, e.g., Newey (1997) and Chen (2007). As stated earlier, this condition is often directly connected to smoothness of the underlying functions.

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

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

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

$$\|\alpha_0\| \leqslant B, \ b \leqslant \mathrm{E}[\zeta_i^2 \mid x_i, v_i], \quad \mathrm{E}[|\zeta_i^q| \mid x_i, v_i] \leqslant B, \quad b \leqslant \mathrm{E}[v_i^2 \mid x_i], \quad \mathrm{E}[|v_i^q| \mid x_i] \leqslant B. \tag{5.32}$$

Let $\mathbf{P}_n$ be the collection of all regression models P that obey the conditions set forth above for a given n. Then, as established in Appendix F, 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 in those conditions depending only on the constants $(\underline{f}, \bar{f}, a, A, b, B, q)$. 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 for any $\mathrm{P} \in \mathbf{P}_n$. In particular, as a special case, the same conclusion applies uniformly in $\mathrm{P} \in \mathbf{P} = \cap_{n \geqslant n_0} \mathbf{P}_n$. □

6. Monte-Carlo Examples

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

All of the simulation results are based on the structural model

$$y_i = d_i'\alpha_0 + x_i'\theta_g + \sigma_y(d_i, x_i)\zeta_i, \quad \zeta_i \sim N(0, 1) \tag{6.33}$$

where $p = \dim(x_i) = 200$, 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. In each design, we generate

$$d_i = x_i'\theta_m + \sigma_d(x_i)v_i, \quad v_i \sim N(0, 1) \tag{6.34}$$

with $\mathrm{E}[\zeta_i v_i] = 0$. 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). Another option would be to use the standard error estimator recently proposed in Cattaneo, Jansson, and Newey (2010).

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 $\sigma_y = \sigma_d = 1$. The second dgp, "Design 2," is heteroscedastic with $\sigma_{d,i} = \sqrt{\frac{(1+x_i'\beta_0)^2}{\mathbb{E}_n(1+x_i'\beta_0)^2}}$ and $\sigma_{y,i} = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\mathbb{E}_n(1+\alpha_0 d_i + x_i'\beta_0)^2}}$. 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 $R^2 = 0, .2, .4, .6$, and $.8$. In the heteroscedastic design, we choose c_y and c_d based on R^2 as if (6.33) and (6.34) 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 $j \leq 5$ and $\theta_{m,j} = c_d(1/j)^2$ for $j \leq 5$. 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 $R^2 = 0, .2, .4, .6$, and $.8$ in the case that all of the random coefficients were 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. In a working paper version of this paper Belloni, Chernozhukov, and Hansen (2011b), 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.[10]

We report results for five different procedures. Two of the procedures are infeasible benchmarks: Oracle and Double-Selection Oracle estimators, which use of knowledge of the true

---

[10] In particular, the post-double-Lasso performed very well across all simulations 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, in no design was the post-double-Lasso outperformed by other feasible procedures. In extensive initial simulations, we also found that Square-Root Lasso and Iterated Lasso performed very similarly and thus only report Lasso results.

coefficient structures θ_g and θ_m and are thus unavailable in practice. The Oracle estimator is the ordinary least squares of $y_i - x_i'\theta_g$ on d_i, and the Double-Selection Oracle is the ordinary least squares of $y - x_i'\theta_g$ on $d_i - x_i'\theta_m$. The other procedures we consider are feasible. In all of them, we rely on Lasso and set λ according to the algorithm outlined in Appendix A with $1 - \gamma = .95$. One procedure is the standard post-single selection estimator – the Post-Lasso – which applies Lasso to equation (6.33) without penalizing α, the coefficient on d, to select additional control variables from among x. Estimates of α_0 are then obtained by OLS regression of y on d 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 on x to select a set of predictors for y and run Lasso of d on x to select a set of predictors for d. α_0 is then estimated by running OLS regression of y on d 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. Post-Double-Selection + Ridge is an *ad hoc* variant of Post-Double-Selection in which we add the ridge fit from equation (6.34) as an additional potential regressor that may be selected by Lasso. The ridge fit is obtained with a single ridge penalty parameter that is chosen using 10-fold cross-validation. This procedure is motivated by a desire to add further robustness in the case that many small coefficients are suspected. 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 $(R_y^2, R_d^2) = (0, .2), (0, .8), (.8, .2)$, and $(.8, .8)$ 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 1, 2, and 3 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 1, 2, and 3 respectively report the results from Design 1, 2, and 3. The figures are plotted with the same scale to aid comparability and for readability rejection frequencies for Post-Lasso were censored at .5. 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 is quite non-uniform across R^2 combinations and does not reliably control size distortions or bias except in the case where the controls are 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 1 and 2, 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 3, 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 4, we compare the performance of the Post-Double-Selection procedure to the *ad hoc* Post-Double-Selection procedure which selects among the original set of variables augmented with the ridge fit obtained from equation (6.34). 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 favorable 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 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 can not be a reliable tool to these researchers.

7. 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 the context 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 the additional discussions in Foote and Goetz (2008) and Donohue III and Levitt (2008) and then present estimates obtained using the methods developed in this paper.

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

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

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

$$y_{cit} = \alpha_c a_{cit} + w_{it}'\beta_c + \delta_{c,i} + \gamma_{c,t} + \varepsilon_{cit} \tag{7.35}$$

where i indexes states, t indexes times, $c \in \{\text{violent, property, murder}\}$ indexes type of crime, $\delta_{c,i}$ are state-specific effects that control for any time-invariant state-specific characteristics, $\gamma_{c,t}$ are time-specific effects that control flexibly for any aggregate trends, w_{it} are a set of control variables to control for time-varying confounding state-level factors, a_{cit} is a measure of the abortion rate relevant for type of crime c,[11] and y_{cit} is the crime-rate for crime type c. Throughout the remainder of this section, we drop the c subscript for convenience but note that separate models are estimated for each crime type and thus all coefficients are allowed to freely vary across crime type. 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 a state having a 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 (7.35) 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 (7.35). We refer the reader to the original paper for additional details, data definitions, and institutional background.

For our analysis, we follow Donohue III and Levitt (2001) and rely on the argument that the abortion rates defined above may be taken as exogenous relative to crime rates conditional upon a set of factors. Unlike Donohue III and Levitt (2001), we do not assume that the identity of these factors is known and allow for smooth, flexible trends to account for unobservable factors that may influence both abortion and crime but smoothly trend over time. Given the seemingly obvious importance of controlling for state and time effects, we account for these in all models we estimate by including a full set of state and time dummies. Thus, we estimate

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

models of the form

$$y_{it} = \alpha a_{it} + w_{it}'\beta_y + \delta_{y,i} + \gamma_{y,t} + g(z_{it}, t) + \zeta_{it} \tag{7.36}$$

$$a_{it} = w_{it}'\beta_a + \delta_{a,i} + \gamma_{a,t} + m(z_{it}, t) + v_{it} \tag{7.37}$$

where $g(z, t)$ and $m(z, t)$ are smooth functions of observed variables z_{it} which includes w_{it}, time-invariant characteristics of $\{y_{it}, a_{it}, w_{it}\}_{t=1}^T$ such as initial conditions or state-level averages, and time. 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 13 time series observations for a total of 624 observations. With these deletions, our baseline estimates using the same controls as in (7.35) 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 of (7.35) are given in the first and second row of Panel A of Table 2.[12]

Note that interpreting estimates of the effect of abortion from model (7.35) 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. Allowing for such variables is important in that one might believe 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-level 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. In practice, it is common to account for this possibility by allowing state-specific trends (e.g. by specifying $g(z_{it}, t) = \kappa_{g,i}t$) in addition to state-specific intercepts. Results from estimating (7.35) with state-specific trends are given in the third row in Table 2 Panel A. In this example, the inclusion of state-specific linear trends renders the results very imprecise. Of course, one might argue that including state-specific linear trends is too aggressive in a sample with only 13 time series observations. The linear trend specification is also very restrictive in imposing that any unobserved factors that relate to both abortion and crime exhibit constant growth over the 13 year time period. The assumption of constant growth becomes even more problematic when one expands the time period as in Foote and Goetz (2008) and Donohue III and Levitt (2008) discussed below.

We follow the Chamberlain (1985) type approach and approximate $g(z_{it}, t)$ and $m(z_{it}, t)$ by a large number of controls. We approximate these functions by forming 27 factors to include

[12]Our estimates differ for three reasons. First, we delete Alaska, Hawaii, and Washington, D.C. Second, Donohue III and Levitt (2001) use population weighted estimates. Third, Donohue III and Levitt (2001) use an FGLS estimator based on an AR(1) model in the errors where the errors across states share the same AR coefficient.

in z_{it},

$$z_{it} = (a_{i0},\ \frac{1}{T}\sum_t a_{it},\ y_{i0},\ w'_{i0},\ \frac{1}{T}\sum_t w'_{it},\ w'_{it})',$$

forming nine smooth function of time,

$$f_t = (t,\ t^2,\ t^3,\ \sin(\pi\frac{t}{T}),\ \sin(2\pi\frac{t}{T}),\ \sin(3\pi\frac{t}{T}),\ \cos(\pi\frac{t}{T}),\ \cos(2\pi\frac{t}{T}),\ \cos(3\pi\frac{t}{T}))',$$

and then supposing that

$$g(z_{it}, t) \approx \sum_{r=1}^{27}\sum_{s=1}^{9} \beta_{g,r,s} z_{it,r} f_{t,s} = h'_{it}\beta_g \qquad \text{and}$$

$$m(z_{it}, t) \approx \sum_{r=1}^{27}\sum_{s=1}^{9} \beta_{m,r,s} z_{it,r} f_{t,s} = h'_{it}\beta_m$$

where h_{it} is the vector containing all the interactions, and β_g and β_m are the vectors of coefficients for each equation. That is, we add an additional 243 control variables to the model and use the methods developed in this paper to search among these 243 additional control variables to see if there are potentially important factors that are missed in equation (7.35).[13] With this set of controls, the models we estimate are all more general than (7.35) and are neither more nor less general than a model with state-specific trends in that we allow for nonlinearity in trends but do not allow for arbitrarily different state-specific coefficients. Rather, we restrict these coefficients to differ depending on values of observable covariates.

Controlling for a large set of variables as described above is desirable from the standpoint of making the belief underlying the causal interpretation of the abortion coefficient, that the abortion rate defined above may be taken as being as good as randomly assigned once the set of variables considered is controlled for, more plausible. As with the inclusion of state-specific trends, the downside is that controlling for many variables lessens our ability to identify the effect of interest and thus tends to make estimates far less precise. For example, the estimated abortion effects conditioning on the full set of 68 variables in (7.35) plus the 243 approximating functions (for a total of 311 control variables) are given in the fourth row of Table 2 Panel A. As expected, all coefficients are estimated very imprecisely. Of course, very few researchers would consider using 311 controls with only 624 observations due to exactly this issue.

We are faced with a trade-off between the precision of the estimate and the plausibility of the conditional exogeneity assumption. By including additional controls in the specification, we make the conditional exogeneity assumption more plausible. At the same time, we potentially reduce the precision of our estimate. The double selection method proposed in this paper

[13] To allow time effects, state effects, and w_{it} to enter each equation without shrinkage, we use our methods based on $\tilde{y}_{it}$, $\tilde{a}_{it}$ and $\tilde{h}_{it}$ where $\tilde{y}_{it}$ is the residual from the regression of y_{it} on w_{it} and a full set of state and time dummies and $\tilde{a}_{it}$ and $\tilde{h}_{it}$ are defined similarly.

offers one rigorous approach to achieving a balance. Thus, the approach complements the usual careful specification analysis by providing a researcher a simple-to-implement, data-driven way to search for a set of influential confounds from among a sensibly chosen broader set of potential confounding variables.

In the abortion example, we use the post-double-Lasso estimator defined in Section 2.2 for each of our dependent variables. For violent crime, a total of 15 variables are selected: eight in the abortion equation[14] and seven in the crime equation.[15] For property crime, 16 variables are selected: ten in the abortion equation[16] and seven in the crime equation[17] with one occurring in both. For murder, ten variables are selected: eight in the abortion equation[18] and two in the crime equation.[19] It is interesting in looking at the selected variables that in all cases initial or average levels of abortion interacted with nonlinear trend terms and initial levels of crime interacted with nonlinear trend terms are selected. This selection illustrates the potential importance of allowing for nonlinear trends and also the potential that there may be omitted factors that are related to both abortion and crime.

Estimates of the causal effect of abortion on crime obtained by searching for confounding factors among our set of 243 potential controls are given in the fifth row of Panel A of Table 2. Each of these estimates is obtained from the least squares regression of the crime rate on the abortion rate, a full set of state dummies, a full set of time dummies, the initial eight controls that vary across states and time from (7.35) and the 15, 16, and ten controls selected by the post-double-Lasso procedure for violent crime, property crime, and murder respectively. The estimates for the effect of abortion on violent crime and the effect of abortion on murder are quite imprecise, producing 95% confidence intervals that encompass large positive and negative values. The estimated effect for property crime is roughly in line with the previous

[14]The selected variables are average abortion times t, average abortion times $\cos(\pi\frac{t}{T})$, initial crime times t^2, initial crime times $\cos(2\pi\frac{t}{T})$, average income times t^3, average income times $\sin(\pi\frac{t}{T})$, average income times $\cos(2\pi\frac{t}{T})$, and initial poverty times $\cos(2\pi\frac{t}{T})$.

[15]The selected variables are average abortion times t^3, initial abortion times t^3, initial abortion times $\sin(\pi\frac{t}{T})$, initial poverty times $\sin(2\pi\frac{t}{T})$, initial poverty times $\cos(\pi\frac{t}{T})$, police_{it} times t^3, and beer_{it} times $\sin(3\pi\frac{t}{T})$.

[16]The selected variables are average abortion times $\cos(\pi\frac{t}{T})$, initial abortion times $\sin(3\pi\frac{t}{T})$, initial crime times $\cos(\pi\frac{t}{T})$, average income times t, average income times $\cos(\pi\frac{t}{T})$, initial poverty times $\cos(2\pi\frac{t}{T})$, initial beer times $\cos(2\pi\frac{t}{T})$, prison_{it} times $\cos(\pi\frac{t}{T})$, income_{it} times $\cos(\pi\frac{t}{T})$, and AFDC_{it} times $\cos(2\pi\frac{t}{T})$.

[17]The selected variables are average abortion times t^3, initial crime times $\sin(2\pi\frac{t}{T})$, initial crime times $\cos(\pi\frac{t}{T})$, average police times $\cos(2\pi\frac{t}{T})$, average AFDC times t, initial AFDC times t, and initial AFDC times t^2.

[18]The selected variables are average abortion times t^2, average abortion times $\cos(\pi\frac{t}{T})$, initial crime times t^3, initial crime times $\cos(2\pi\frac{t}{T})$, average income times t^3, average income times $\sin(\pi\frac{t}{T})$, average income times $\cos(2\pi\frac{t}{T})$, and average income times $\cos(3\pi\frac{t}{T})$.

[19]The variables selected are average abortion times $\sin(\pi\frac{t}{T})$ and initial abortion times $\sin(\pi\frac{t}{T})$.

estimates though it is no longer significant and has a 95% confidence interval that includes negative as well as modest positive effects. 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 .2522, .3533, and .0554 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 .9906, .9039, and .9863 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 post-double-selection procedure performed quite well, while the standard post-single-selection procedures performed quite poorly.

While the inclusion of trigonometric terms in our approximations allows for capturing some types of cyclicality, some researchers may feel more comfortable restricting attention to simpler trend specifications. To allow for this, we also present results in which the trigonometric functions are dropped from f_t, so that

$$f_t = (t,\ t^2,\ t^3).$$

That is, we approximate the functions as $g(z_{it}, t) \approx \sum_{r=1}^{27} \sum_{s=1}^{3} \beta_{g,r,s} z_{it,r} f_{t,s} = h_{it}' \beta_g$ and $m(z_{it}, t) \approx \sum_{r=1}^{27} \sum_{s=1}^{3} \beta_{m,r,s} z_{it,r} f_{t,s} = h_{it}' \beta_m$ which allows only cubic polynomial trends interacted with state-level characteristics. In this case, only 81 terms are considered in addition to the 68 controls from the original specification. Results using all 149 controls are given in the row "Polynomial Trends" in Table 2 Panel A, and results based on Lasso selection among the 81 added controls are given in the row "Post-Double-Selection, Polynomial Trends." Looking at these results we see that we would draw the same qualitative conclusion using this restricted specification as we would when allowing for trigonometric terms as well. Specifically, the estimated abortion effects become quite imprecise after allowing only for the polynomial terms in time.[20]

A similar conclusion was reached by Foote and Goetz (2008) who, without doing formal variable selection, found that inclusion of a linear trend interacted with the average crime rate from a period before the abortion rate should have been able to have an effect on the crime rate substantially attenuated the estimated effects from Donohue III and Levitt (2001) and also rendered them imprecise. It is interesting that we reach a similar conclusion through the use of formal variable selection procedures motivated by the desire to allow for allow flexible,

[20] In addition to the 68 original variables, the double-selection procedure selects ten total additional variables for the violent crime regression, eight additional variables for the property crime regression, and five additional variables for the murder regression. In each case, the mean of the abortion rate times t is selected and this variable accounts for most of the explanatory power among the selected additional regressors.

yet parsimonious trends in an effort to make the exogeneity assumption conditional on controls more plausible.

In a response to Foote and Goetz (2008), Donohue III and Levitt (2008) note that one problem with allowing flexible trends is that the short time series renders estimates of the treatment effect imprecise once flexible trends are allowed. Specifically, estimated treatment effects are imprecise in their preferred specification

$$y_{it} = \alpha a_{it} + \delta_i + \gamma_{d,t} + \kappa_i t + \varepsilon_{it} \tag{7.38}$$

where δ_i is a state-specific effect, κ_i is a state-specific coefficient on a linear trend, and $\gamma_{d,t}$ is Census division $\times$ time effect. To address this issue, Donohue III and Levitt (2008) extend the sample period to 1960-2003 to allow more precise estimates of the trends and thus more reliable estimates of the treatment effect. They find that the results in this longer sample with the full set of division times time interactions and state-specific trends are similar to the initial results in the shorter panel. Results from this analysis in Donohue III and Levitt (2008) are provided in the first row of Panel B of Table 2. In the second row of Table 2, Panel B, we report results from our estimates of the abortion effect using data from 1960-2003 using exactly the same methodology as Donohue III and Levitt (2008), and we report results from simple OLS regression of (7.38) in the third row.[21]

While (7.38) is certainly more general than (7.35), state-specific linear trends are still quite restrictive, especially over a time period of 40 years. Specifically, it is a strong assumption that unobserved factors that are correlated to both state level abortion and crime rates exhibited constant growth over such a long time period. To allow for smooth, but flexible trends, we once again consider variable selection in a more general model

$$y_{it} = \alpha a_{it} + \delta_{y,i} + \gamma_{y,d,t} + \kappa_{y,i} t + g(z_{it}, t) + \zeta_{it} \tag{7.39}$$

$$a_{it} = \delta_{a,i} + \gamma_{a,d,t} + \kappa_{d,i} t + m(z_{it}, t) + v_{it} \tag{7.40}$$

where $g(z,t)$ and $m(z,t)$ are smooth functions of observed variables z_{it} which includes time-invariant characteristics of $\{y_{it}, a_{it}, w_{it}\}_{t=1}^{T}$ such as initial conditions or state-level averages and

[21]Our results differ due to the exclusion of Alaska, Hawaii, and Washington, D.C. We also completed the data on abortion before 1985 by filling in 0 for all abortion rates before 1985.

time. For this longer time period, we approximate g and m by setting

$$z_{it} = (a_{i1985},\ \frac{1}{44}\sum_{t=1960}^{2003} a_{it},\ y_{i1960},\ y_{i1961},\ w'_{i1985},\ \frac{1}{13}\sum_{t=1985}^{1997} w'_{it})',$$

$$f_t = (t^2,\ t^3,\ t^4,\ t^5,\ \sin(\pi\frac{t}{T}),\ \sin(2\pi\frac{t}{T}),\ \sin(3\pi\frac{t}{T}),\ \sin(4\pi\frac{t}{T}),$$
$$\cos(\pi\frac{t}{T}),\ \cos(2\pi\frac{t}{T}),\ \cos(3\pi\frac{t}{T}),\ \cos(4\pi\frac{t}{T}))',$$

and then supposing

$$g(z_{it}, t) \approx \sum_{r=1}^{20}\sum_{s=1}^{12} \beta_{g,r,s} z_{it,r} f_{t,s} = h'_{it}\beta_g \quad \text{and}$$
$$m(z_{it}, t) \approx \sum_{r=1}^{20}\sum_{s=1}^{12} \beta_{m,r,s} z_{it,r} f_{t,s} = h'_{it}\beta_m,$$

where h_{it} is the vector containing all the interactions, and β_g and β_m are the vectors of coefficients for each equation. Thus, we add an additional 240 control variables to (7.38).[22]

Estimates of the abortion effect using the full set of 713 controls consisting of the 473 controls in (7.38) augmented with the 240 additional controls for smooth nonlinear trends are given in the fourth row of Table 2 Panel B. As expected, the estimated abortion effects are extremely imprecise given this large set of controls.

To pare down the number of controls, we employ the Double-Selection procedure developed in this paper to search for a smaller set of relevant controls among the 240 potential additions. Based on this exercise, we select a total of 31 additional variables for the violence equation, 30 for the abortion equation, and 27 for the murder equation. R^2's from the regression of crime rates on the controls are .2806, .3451, and .0422 for violent crime, property crime, and the murder rate respectively; and the R^2's from regressing the abortion rate on the selected variables are .9618, .9461, and .9775 for violent crime, property crime, and the murder rate respectively. Estimates of the treatment effect controlling for the variables in Donohue III and Levitt (2008) and those selected by Double-Selection are given in the "Post-Double-Selection" row of Table 2, Panel B. As in the original data, we find that estimates of the abortion effect are relatively imprecise once parsimonious nonlinear trends are allowed for.

As in the previous specification, we report results using only interactions with the polynomial trend terms, i.e.

$$f_t = (\ t^2,\ t^3,\ t^4,\ t^5)',$$

[22]To allow for all the effects in (7.38) to enter each equation without shrinkage, we use our methods based on $\tilde{y}_{it}$, $\tilde{a}_{it}$ and $\tilde{h}_{it}$ where $\tilde{y}_{it}$ is the residual from the regression of y_{it} on a full set of state dummies, a full set of Census division cross time dummies, and a full set of state-specific trends and $\tilde{a}_{it}$ and $\tilde{h}_{it}$ are defined similarly.

in the final two rows of Panel B of Table 2. Using only the interactions with the polynomial terms adds 80 potential regressors to the 473 included in the original Donohue III and Levitt (2008) specification. Results using the full set of 553 regressors are reported in the row "Polynomial Trends" in Table 2 Panel B and show that once again using this broad set of regressors results in imprecise estimates of the regression coefficients. The lack of precision in the estimated abortion effect is qualitatively unchanged after using the double-selection procedure to select controls from among this restricted set, again illustrating that the baseline result is not driven by the inclusion of trigonometric terms in the set of approximating functions.[23]

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. In this example, we see that inference about the treatment effects using the variable selection method differs substantively from inference drawn using the original set of controls. This statement is true whether one considers the data and model from Donohue III and Levitt (2001) or Donohue III and Levitt (2008). This difference is driven by the variable selection method's selecting different variables than are usually considered. Thus, it appears that the usual interpretation of there being a substantive causal effect of abortion on crime hinges on strong prior beliefs about the types of trends that may appear in the structural equation. In particular, inclusion of a modest number of smooth nonlinear trends interacted with time-invariant state-level characteristics substantively increases the variance of the estimated treatment effects.

8. Conclusion

In this paper, we consider estimation of treatment effects or structural parameters in an environment where the treatment is believed to be exogenous conditional on observables. We do not impose the conventional assumption that the identities of the relevant conditioning variables and the functional form with which they enter the model are known. Rather, we assume that the researcher believes there is a relatively small number of important factors whose identities are unknown within a much larger known set of potential variables and transformations. This sparsity assumption allows the researcher to estimate the desired treatment effect and

[23]In addition to the 473 original variables, the double-selection procedure selects 12 total additional variables for the violent crime regression, 11 additional variables for the property crime regression, and 11 additional variables for the murder regression.

infer a set of important variables upon which one needs to condition by using modern variable selection techniques without *ex ante* knowledge of which are the important conditioning variables. Since naive application of variable selection methods in this context may result in very poor properties for inferring the treatment effect of interest, we propose a "double-selection" estimator of the treatment effect, provide a formal demonstration of its properties for estimating the treatment effect, and provide its approximate distribution under technical regularity conditions and the assumed sparsity in the model.

In addition to the theoretical development, we illustrate the potential usefulness of our proposal through a number of simulation studies and an empirical example. In Monte Carlo simulations, our procedure outperforms simple variable selection strategies for estimating the treatment effect across the designs considered and does relatively well compared to an infeasible estimator that uses the identities of the relevant conditioning variables. We then apply our estimator to attempt to estimate the causal impact of abortion on crime following Donohue III and Levitt (2001). We find that our procedure selects a small number of conditioning variables. After conditioning on these selected variables, one would draw qualitatively different inference about the effect of abortion on crime than would be drawn if one assumed that the correct set of conditioning variables was known and the same as those variables used in Donohue III and Levitt (2001). Taken together, the empirical and simulation examples demonstrate that the proposed method may provide a useful complement to other sorts of specification analysis done in applied research.

Appendix A. Iterated Estimation of Penalty Loadings

In the case of Lasso under heteroscedasticity, we must specify for the penalty loadings (2.12). Here we state algorithms for estimating these loadings.

Let I_0 be an initial set of regressors with bounded number of elements, including for example intercept. Let $\bar\beta(I_0)$ be the least squares estimator of the coefficients on the covariates associated with I_0, and define $\widehat l_{j0} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\bar\beta(I_0))^2]}$.

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* $l_{j,k+1} := \sqrt{\mathbb{E}_n[x_{ij}^2(y_i - x_i'\widehat\beta)^2]}\sqrt{n/(n-\widehat s)}$. *(3) If* $\max_{1\leqslant j\leqslant p} |\widehat l_{j,k} - \widehat l_{j,k+1}| \leqslant \nu$ *or* $k > K$, *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).*

A similar algorithm can be defined for using with Lasso instead of Post-Lasso.

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

A similar algorithm can be defined for using with Square-Root-Lasso instead of Post-Square-Root-Lasso.

Appendix B. Proof of Theorem 1

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

Let $Y=[y_1,...,y_n]'$, $X=[x_1,...,x_n]'$, $D=[d_1,...,d_n]'$, $V=[v_1,...,v_n]'$, $\zeta=[\zeta_1,...,\zeta_n]'$, $m=[m_1,...,m_n]'$, $R_m=[r_{m1},...,r_{mn}]'$, $g=[g_1,...,g_n]'$, $R_g=[r_{g1},...,r_{gn}]'$, and so on. For $A\subset\{1,...,p\}$, let $X[A]=\{X_j, j\in A\}$, where $\{X_j, j=1,...,p\}$ are the columns of X. Let

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

be the projection operator sending vectors in $\mathbb{R}^n$ onto 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 $Z\in\mathbb{R}^n$, let

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

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

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

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

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

By Steps 2 and 3,

$$ii = V'V/n + o_P(1) \text{ and } i = V'\zeta/\sqrt{n} + o_P(1).$$

Next note that $V'V/n = \mathrm{E}[V'V/n] + o_P(1)$ by Chebyshev, and because $\mathrm{E}[V'V/n]$ 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 $\sigma_n^2 = \bar{\mathrm{E}}[v_i^2]^{-1}\bar{\mathrm{E}}[\zeta_i^2 v_i^2]\bar{\mathrm{E}}[v_i^2]^{-1}$ 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 $z_{i,n} := \sigma_n^{-1} v_i \zeta_i$ are i.n.i.d. with mean zero. For $\delta > 0$ such that $4 + 2\delta \leqslant q$

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

by Condition SM. This condition verifies the Lyapunov condition and thus 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 $D = m + V$,

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

First, by Step 5 and 6 below we have

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

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

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

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

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

Step 3. (Behavior of ii.) Decompose

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

Then $|ii_a| \lesssim_P [s\log(p\vee n)]/n = o_P(1)$ by Step 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

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

To show (a), we use Lemma 4 stated in Appendix F 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 $1 - o(1)$

$$\max_{1\leqslant j\leqslant p}\left|\frac{n^{-1/2}\sum_{i=1}^n x_{ij}\zeta_i}{\sqrt{\mathbb{E}_n[x_{ij}^2\zeta_i^2]}}\right| \leqslant \Phi^{-1}\left(1-\frac{1}{2\ell_n p}\right) \lesssim \sqrt{2\log(2\ell_n p)} \lesssim \sqrt{\log(p\vee n)}. \tag{B.41}$$

By Lemma 4 the first inequality in (B.41) holds, provided that for all n sufficiently large the following holds,

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

Since we can choose ℓ_n to grow as slowly as needed, a sufficient condition for this are the conditions:

$$\log p = o(n^{1/3}) \text{ and } \min_{1\leqslant j\leqslant p} M_j \gtrsim 1,$$

which both hold by Condition SM. Finally,

$$\max_{1\leqslant j\leqslant p}\mathbb{E}_n[x_{ij}^2\zeta_i^2] \lesssim_{\mathrm{P}} 1, \tag{B.42}$$

by Condition SM. Therefore (a) follows from the bounds (B.41) and (B.42). 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

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

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 $\|\mathcal{M}_{\widehat{I}_2}(\alpha_0 m + g)/\sqrt{n}\| \leqslant \|(X\tilde{\beta}_{Y_1}(\widehat{I}_2) - \alpha_0 m - g)/\sqrt{n}\| \lesssim_P \sqrt{[s\log(p\vee n)]/n}$, (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

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

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

$$\mathbb{E}_n[\widehat v_i^2\widehat\zeta_i^2] - \bar{\mathrm{E}}[v_i^2\zeta_i^2] \to_P 0, \quad \mathbb{E}_n[\widehat v_i^2] - \bar{\mathrm{E}}[v_i^2] \to_P 0,$$

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 triangular 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

$$iii_a \leqslant 2\max_{i\leqslant n} d_i^2|\alpha_0-\check\alpha|^2\mathbb{E}_n[\widehat v_i^2] \lesssim_P n^{(2/q)-1} = o(1) \quad \text{(B.43)}$$

$$iii_c \leqslant 2\max_{i\leqslant n}\{|\tilde\zeta_i||d_i|\}\mathbb{E}_n[\widehat v_i^2]|\alpha_0-\check\alpha| \lesssim_P n^{(2/q)-(1/2)} = o(1) \quad \text{(B.44)}$$

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

$$\max_{i\leqslant n}|d_i| + |\tilde\zeta_i| + |\tilde v_i| \lesssim_P n^{1/q}. \quad \text{(B.45)}$$

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 (B.43)-(B.44).

Next we bound,

$$\begin{aligned} iii_d &\leqslant 2\max_{i\leqslant n}|\tilde{\zeta}_i|\max_{i\leqslant n}|x_i'(\check{\beta}-\beta_{g0})|\mathbb{E}_n[\widehat{v}_i^2] \\ &\lesssim_P n^{1/q}\max_{i\leqslant n}\|x_i\|_\infty\sqrt{\frac{s}{\sqrt{n}}\frac{s\log(p\vee n)}{\sqrt{n}}} = o_P(1), \end{aligned} \tag{B.46}$$

using (B.45) and that for $\widehat{T}_g = \text{support}(\beta_{g0})\cup\widehat{I}$, we have

$$\max_{i\leqslant n}\{x_i'(\check{\beta}-\beta_{g0})\}^2 \leqslant \max_{i\leqslant n}\|x_{i\widehat{T}_g}\|^2\|\check{\beta}-\beta_{g0}\|^2,$$

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, and since $\check{\beta}[\widehat{I}] = (X[\widehat{I}]'X[\widehat{I}])^{-}X[\widehat{I}]'(\zeta+g-(\check{\alpha}-\alpha_0)D)$ we have

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

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 (B.46) then follows by condition ASTE (iv) and (v).

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

$$\begin{aligned} iii_b &\leqslant 2\max_{i\leqslant n}\{x_i'(\check{\beta}-\beta_{g0})\}^2\mathbb{E}_n[\widehat{v}_i^2] \\ &\lesssim_P \max_{i\leqslant n}\|x_i\|_\infty^2\frac{s}{\sqrt{n}}\frac{s\log(p\vee n)}{\sqrt{n}} = o_P(1). \end{aligned} \tag{B.47}$$

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

Appendix C. 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 $n \in \{n_0, n_0+1, ...\}$, with $\mathrm{P}_n \in \mathbf{P}_n$ for each $n \in \{n_0, n_0+1, ...\}$. 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 $k \in \{1, 2, ...\}$

$$\lim_{k\to\infty} \mathrm{P}_{n_k}\left(\alpha_0 \in [\check{\alpha} \pm c\widehat{\sigma}_{n_k}/\sqrt{n_k}]\right) = 1-\gamma. \tag{C.48}$$

Suppose that the claim of corollary does not hold, i.e.

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

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 (C.48). The claim (i) follows. Claim (ii) follows from claim (i), since $\mathbf{P} \subseteq \mathbf{P}_n$ for all $n \geqslant n_0$. □

Appendix D. 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 $D = m + V$

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

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

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

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

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

Then $|ii_a| \lesssim_P o(n^{1/2})/n = o_P(n^{-1/2})$ 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

$$\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\| \geqslant |\ \|X(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})/\sqrt{n}\| - \|R_g/\sqrt{n}\|\ |$$

where $\|R_g/\sqrt{n}\| \lesssim_P \sqrt{s/n}$ holds by Chebyshev inequality and Condition ASTE(iii). Moreover, by Condition HLMS we have $\|\mathcal{M}_{\widehat{I}}g/\sqrt{n}\| = o_P(n^{-1/4})$ and $\widehat{s} = |\widehat{I}| \lesssim_P s$. Thus

$$\begin{aligned} o(n^{-1/4}) + \sqrt{s/n} &\gtrsim_P \|X(\tilde{\beta}_g(\widehat{I}) - \beta_{g0})/\sqrt{n}\| \\ &\geqslant \sqrt{\phi_{\min}(s+\widehat{s})}\|\tilde{\beta}_g(\widehat{I}) - \beta_{g0}\| \\ &\gtrsim_P \|\tilde{\beta}_g(\widehat{I}) - \beta_{g0}\| \end{aligned}$$

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.

□

Appendix E. Proof of Corollary 2

The proof is similar to the proof of Corollary 1.

Appendix F. Verification of Conditions for the Examples

F.1. **Verification for Example 1.** Let $\mathbf{P}$ be the collection of all regression models P that obey the conditions set forth above for all n for the given constants (p, b, B, q_x, q). Below we provide explicit bounds for κ', κ'', c, C, δ_n and Δ_n that appear in Conditions ASTE, SE and SM that depend only on (p, b, B, q_x, q) and n which in turn establish these conditions for any $\mathrm{P} \in \mathbf{P}$.

Condition ASTE(i) is assumed. Condition ASTE(ii) holds with $\|\alpha_0\| \leqslant C_1^{ASTE} = B$. Condition ASTE(iii) holds with $s = p$ and $r_{gi} = r_{mi} = 0$.

Condition ASTE(iv) holds with $\delta_{1n}^{ASTE} := p^2 \log^2(p \vee n)/n \to 0$ since $s = p$ is fixed. Finally, we verify ASTE(v). Because $\tilde{v}_i = v_i$, $\tilde{\zeta}_i = \zeta_i$ and the moment condition $\mathrm{E}[|v_i^q|] + \mathrm{E}[|\zeta_i^q|] \leqslant C_2^{ASTE} = 2B$ with $q > 4$, the first two requirements follow. To show the last requirement, note that because $\mathrm{E}[\|x_i\|^{q_x}] \leqslant B$ we have

$$\mathrm{P}\left(\max_{1\leqslant i\leqslant n} \|x_i\|_\infty > t_{1n}\right) \leqslant \mathrm{P}\left(\left[\sum_{i=1}^n \|x_i\|^{q_x}\right]^{1/q_x} > t_{1n}\right) \leqslant n\mathrm{E}[\|x_i\|^{q_x}]/t_{1n}^{q_x} \leqslant nB/t_{1n}^{q_x} =: \Delta_{1n}^{ASTE}. \tag{F.49}$$

Let $t_{1n} = (n\log n)^{1/q_x} B^{1/q_x}$ so that $\Delta_{1n}^{ASTE} = 1/\log n$. Thus we have with probability $1 - \Delta_{1n}^{ASTE}$

$$\max_{1\leqslant i\leqslant n} \|x_i\|_\infty^2 s n^{-1/2+2/q} \leqslant (n\log n)^{2/q_x} B^{2/q_x} p n^{-1/2+2/q} =: \delta_{2n}^{ASTE}.$$

It follows that $\delta_{2n}^{ASTE} \to 0$ by the assumption that $4/q_x + 4/q < 1$.

To verify Condition SE note that

$$\begin{aligned} P(\|\mathbb{E}_n[x_i x_i'] - E[x_i x_i']\| > t_{2n}) &\leqslant \sum_{k=1}^{p}\sum_{j=1}^{p} \frac{E[x_{ij}^2 x_{ik}^2]}{n t_{2n}^2} \leqslant \sum_{k=1}^{p}\sum_{j=1}^{p} \frac{\sqrt{E[x_{ij}^4]E[x_{ik}^4]}}{n t_{2n}^2} \\ &\leqslant \sum_{k=1}^{p}\sum_{j=1}^{p} \frac{E[\|x_i\|^4]}{n t_{2n}^2} \leqslant \frac{p^2 B^{4/q_x}}{n t_{2n}^2} =: \Delta_{1n}^{SE}. \end{aligned}$$

Setting $t_{2n} := b/2$ we have $\Delta_{1n}^{SE} = (2/b)^2 B^{4/q_x} p^2/n \to 0$ since p is fixed. Then, with probability $1 - \Delta_{1n}^{SE}$ we have

$$\begin{aligned} \lambda_{\min}(\mathbb{E}_n[x_i x_i']) &\geqslant \lambda_{\min}(E[x_i x_i']) - \|\mathbb{E}_n[x_i x_i'] - E[x_i x_i']\| \geqslant b/2 =: \kappa', \\ \lambda_{\max}(\mathbb{E}_n[x_i x_i']) &\leqslant \lambda_{\max}(E[x_i x_i']) + \|\mathbb{E}_n[x_i x_i'] - E[x_i x_i']\| \leqslant E[\|x_i\|^2] + b/2 \leqslant 2B^{2/q_x} =: \kappa''. \end{aligned}$$

In the verification of Condition SM note that the second and third requirements in Condition SM(i) hold with $c_1^{SM} = b$ and $C_1^{SM} = B^{2/q}$. Condition SM(iii) holds with $\delta_{1n}^{SM} := \log^3 p/n \to 0$ since p is fixed.

The first requirement in Condition SM(i) and Condition SM(ii) hold by the stated moment assumptions, for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$, $\tilde{y}_i = d_i$ and $\tilde{y}_i = y_i$,

$$\begin{aligned} E[|\epsilon_i^q|] &\leqslant B =: A_1 \\ E[|d_i^q|] &\leqslant 2^{q-1}E[|x_i'\beta_{m0}|^q] + 2^{q-1}E[|v_i^q|] \leqslant 2^{q-1}E[\|x_i\|^q]\|\beta_{m0}\|^q + 2^{q-1}E[|v_i^q|] \\ &\leqslant 2^{q-1}(B^{q/q_x}B^q + B) =: A_2 \\ E[d_i^4] &\leqslant 2^3(B^{4/q_x}B^4 + B) =: A_2' \\ E[y_i^4] &\leqslant 3^3\|\alpha_0\|^4 E[d_i^4] + 3^3\|\beta_{g0}\|^4 E[\|x_i\|^4] + 3^3 E[\zeta_i^4] \\ &\leqslant 3^3 B^4 2^3 A_2' + 3^3 B^4 B^{4/q_x} + 3^3 B^{4/q} =: A_3 \\ \max_{1\leqslant j\leqslant p} E[x_{ij}^2\tilde{y}_i^2] &\leqslant \max_{1\leqslant j\leqslant p}(E[x_{ij}^4])^{1/2}(E[\tilde{y}_i^4])^{1/2} \leqslant B^{2/q_x}(E[\tilde{y}_i^4])^{1/2} \leqslant B^{2/q_x}(A_2' \vee A_3)^{1/2} =: A_4 \\ \max_{1\leqslant j\leqslant p} E[|x_{ij}\epsilon_i|^3] &= \max_{1\leqslant j\leqslant p} E[|x_{ij}^3|E[|\epsilon_i^3| \mid x_i]] \leqslant B^{3/q}\max_{1\leqslant j\leqslant p} E[|x_{ij}^3|] \leqslant B^{3/q+3/q_x} =: A_5 \\ \max_{1\leqslant j\leqslant p} 1/E[x_{ij}^2] &\leqslant 1/\lambda_{\min}(E[x_i x_i']) \leqslant 1/b =: A_6 \end{aligned}$$

since $4 < q \leqslant q_x$. Thus these conditions hold with $C_2^{SM} = A_2 \vee (A_1 + (A_2' \vee A_3)^{1/2} + A_4 + A_5 + A_6)$.

Next we show Condition SM(iv). By (F.49) we have $\max_{1\leqslant i\leqslant n}\|x_i\|_\infty^2 \leqslant (n\log n)^{2/q_x}B^{2/q_x}$ with probability $1 - \Delta_{1n}^{ASTE}$, thus with the same probability

$$\max_{i\leqslant n}\|x_i\|_\infty^2 \frac{s\log(n\vee p)}{n} \leqslant (B\log n)^{2/q_x}\frac{n^{2/q_x}p\log(p\vee n)}{n} =: \delta_{1n}^{SM} \to 0$$

since $q_x > 4$ and $s = p$ is fixed.

Next for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$ we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |(\mathbb{E}_n - \mathrm{E})[x_{ij}^2\epsilon_i^2]| > \delta_{2n}^{SM} \right) \leqslant \sum_{j=1}^{p} \frac{\mathrm{E}[x_{ij}^4\epsilon_i^4]}{n(\delta_{2n}^{SM})^2} \leqslant \frac{pB^{4/q+4/q_x}}{n(\delta_{2n}^{SM})^2} =: \Delta_{1n}^{SM}$$

by the union bound, Chebyshev inequality and by $\mathrm{E}[x_{ij}^4\epsilon_i^4] = \mathrm{E}[x_{ij}^4 \mathrm{E}[\epsilon_i^4 \mid x_i]] \leqslant B^{4/q+4/q_x}$. Letting $\delta_{2n}^{SM} = B^{2/q+2/q_x} n^{-1/4} \to 0$ we have $\Delta_{1n}^{SM} = p/n^{1/2} \to 0$ since p, B, q and q_x are fixed.

Next for $\tilde{y}_i = d_i$ and $\tilde{y}_i = y_i$ we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |(\mathbb{E}_n - \mathrm{E})[x_{ij}^2\tilde{y}_i^2]| > \delta_{3n}^{SM} \right) \leqslant \sum_{j=1}^{p} \frac{\mathrm{E}[x_{ij}^4\tilde{y}_i^4]}{n(\delta_{3n}^{SM})^2} \leqslant \frac{pB^{4/q_x}A_8^{4/q}}{n(\delta_{3n}^{SM})^2} =: \Delta_{2n}^{SM}$$

by the union bound, Chebyshev inequality and by

$$\mathrm{E}[x_{ij}^4\tilde{y}_i^4] \leqslant \mathrm{E}[x_{ij}^{\tilde{q}}]^{4/\tilde{q}} \mathrm{E}[\tilde{y}_i^q]^{4/q} \leqslant \mathrm{E}[x_{ij}^{q_x}]^{4/q_x} \mathrm{E}[\tilde{y}_i^q]^{4/q} \leqslant B^{4/q_x} A_8^{4/q}$$

holding by Hölder inequality where $4 < \tilde{q} \leqslant q_x$ such that $4/q + 4/\tilde{q} = 1$, and

$$\begin{aligned} \mathrm{E}[\tilde{y}_i^q] &\leqslant (1 + 3^{q-1}\|\alpha_0\|^q)\mathrm{E}[d_i^q] + 3^{q-1}\|\beta_{g0}\|^q \mathrm{E}[\|x_i\|^q] + 3^{q-1}\mathrm{E}[\zeta_i^q] \\ &\leqslant 3^q(A_2 + B^q A_2 + B^q B^{q/q_x} + B) =: A_8. \end{aligned}$$

Letting $\delta_{3n}^{SM} = B^{4/q_x} A_8^{4/q} n^{-1/4} \to 0$ we have $\Delta_{2n}^{SM} = p/n^{1/2} \to 0$ since p, B, q and q_x are fixed.

Finally, we set $c = c_1^{SM}$, $C = \max\{C_1^{ASTE}, C_2^{ASTE}, C_1^{SM}, C_2^{SM}\}$, $\delta_n = \max\{\delta_{1n}^{ASTE}, \delta_{2n}^{ASTE}, \delta_{1n}^{SM} + \delta_{2n}^{SM} + \delta_{3n}^{SM}\} \to 0$, and $\Delta_n = \max\{\Delta_{1n}^{ASTE} + \Delta_{1n}^{SM} + \Delta_{2n}^{SM}, \Delta_{1n}^{SE}\} \to 0$. $\square$

We will make use of the following technical lemma in the verification of examples 2, 3, and 4.

Lemma 2 (Uniform Approximation)**.** *Let $h_i = x_i'\theta_h + \rho_i$ be a function whose coefficients $\theta_h \in S_A^a(p)$, and $\underline{\kappa} \leqslant \lambda_{\min}(\mathrm{E}[x_i x_i']) \leqslant \lambda_{\max}(\mathrm{E}[x_i x_i']) \leqslant \bar{\kappa}$. For $s = A^{1/a} n^{1/2a}$, $a > 1$, define β_{h0} as in (5.26), $r_{hi} = h_i - x_i'\beta_{h0}$, for $i = 1, \ldots, n$. Then we have*

$$|r_{hi}| \leqslant \|x_i\|_\infty (\bar{\kappa}/\underline{\kappa})^{3/2} \left\{ \frac{2a-1}{a-1} \sqrt{s^2/n} + 5\sqrt{s\mathrm{E}[\rho_i^2]/\underline{\kappa}} \right\} + |\rho_i|.$$

Proof. Let T_h denote the support of β_{h0} and S denote the support of the s largest components of θ_h. Note that $|T_h| = |S| = s$. First we establish some auxiliary bounds on the $\|\theta_h[T_h^c]\|$ and $\|\theta_h[T_h^c]\|_1$. By the optimality of T_h and β_{h0} we have that

$$\sqrt{\mathrm{E}[(h_i - x_i'\beta_{h0})^2]} \leqslant \sqrt{\mathrm{E}[(x_i[S^c]'\theta_h[S^c] + \rho_i)^2]} \leqslant \sqrt{\bar{\kappa}}\|\theta_h[S^c]\| + \sqrt{\mathrm{E}[\rho_i^2]} \text{ and}$$

$$\sqrt{\mathrm{E}[(h_i - x_i'\beta_{h0})^2]} = \sqrt{\mathrm{E}[\{x_i'(\theta_h - \beta_{h0}) + \rho_i\}^2]} \geqslant \sqrt{\underline{\kappa}}\|\theta_h[T_h^c]\| - \sqrt{\mathrm{E}[\rho_i^2]}.$$

Thus we have $\|\theta_h[T_h^c]\| \leqslant \sqrt{\bar\kappa/\underline{\kappa}}\|\theta_h[S^c]\| + 2\sqrt{\mathrm{E}[\rho_i^2]/\underline{\kappa}}$. Moreover, since $\theta_h \in S_A^a(p)$, we have

$$\|\theta_h[S^c]\|^2 = \sum_{j=s+1}^{\infty} \theta_{(j)}^2 \leqslant A^2 \sum_{j=s+1}^{\infty} j^{-2a} \leqslant A^2 s^{-2a+1}/[2a-1] \leqslant A^2 s^{-2a+1}$$

since $a > 1$. Combining these relations we have

$$\begin{aligned} \|\theta_h[T_h^c]\| &\leqslant \sqrt{\bar\kappa/\underline{\kappa}}As^{-a+1/2} + 2\sqrt{\mathrm{E}[\rho_i^2]/\underline{\kappa}} \\ &= \sqrt{\bar\kappa/\underline{\kappa}}\sqrt{s/n} + 2\sqrt{\mathrm{E}[\rho_i^2]/\underline{\kappa}}. \end{aligned}$$

The second bound follows by observing that

$$\begin{aligned} \|\theta_h[T_h^c]\|_1 &\leqslant \sqrt{s}\|\theta_h[T_h^c \cap S]\| + \|\theta_h[S^c]\|_1 \leqslant \sqrt{s}\|\theta_h[T_h^c]\| + As^{-a+1}/[a-1] \\ &\leqslant \sqrt{s^2/n}\sqrt{\bar\kappa/\underline{\kappa}} + 2\sqrt{s\mathrm{E}[\rho_i^2]/\underline{\kappa}} + (s/\sqrt{n})/[a-1] \\ &\leqslant \sqrt{s^2/n}\sqrt{\bar\kappa/\underline{\kappa}}\ a/[a-1] + 2\sqrt{s\mathrm{E}[\rho_i^2]/\underline{\kappa}}. \end{aligned}$$

By the first-order optimality condition of the problem (5.26) that defines β_{h0}, we have

$$\mathrm{E}[x_i[T_h]x_i[T_h]'](\beta_{h0}[T_h] - \theta_h[T_h]) = \mathrm{E}[x_i[T_h]x_i[T_h^c]']\theta_h[T_h^c] + \mathrm{E}[x_i[T_h]\rho_i].$$

Thus, since $\|\mathrm{E}[x_i[T_h]\rho_i]\| = \sup_{\|\eta\|=1}\mathrm{E}[\eta' x_i[T_h]\rho_i] \leqslant \sup_{\|\eta\|=1}\sqrt{\mathrm{E}[(\eta' x_i[T_h])^2]}\sqrt{\mathrm{E}[\rho_i^2]}$ we have

$$\begin{aligned} \underline{\kappa}\|\beta_{h0} - \theta_h[T_h]\| &\leqslant \bar\kappa\|\theta_h[T_h^c]\| + \sqrt{\bar\kappa\mathrm{E}[\rho_i^2]} \\ &\leqslant \bar\kappa^{3/2}As^{-a+1/2}/\sqrt{\underline{\kappa}} + 2\bar\kappa\sqrt{\mathrm{E}[\rho_i^2]/\underline{\kappa}} + \sqrt{\bar\kappa\mathrm{E}[\rho_i^2]} \\ &= \sqrt{s/n}\ (\bar\kappa^{3/2}/\sqrt{\underline{\kappa}}) + \sqrt{\mathrm{E}[\rho_i^2]}\ \sqrt{\bar\kappa}(1 + 2\sqrt{\bar\kappa/\underline{\kappa}}) \end{aligned}$$

where the last inequality follows from the definition of $s = A^{1/a}n^{1/2a}$. Therefore

$$\begin{aligned} |r_{hi}| &= |h_i - x_i'\beta_{h0}| = |x_i'(\theta_h - \beta_{h0})| + |\rho_i| \\ &\leqslant \|x_i\|_\infty\|\theta_h - \beta_{h0}\|_1 + |\rho_i| \\ &\leqslant \sqrt{s}\|x_i\|_\infty\|\theta_{hT_h} - \beta_{h0}\| + \|x_i\|_\infty\|\theta_{hT_h^c}\|_1 + |\rho_i| \\ &\leqslant \|x_i\|_\infty\{\sqrt{s^2/n}\ (\bar\kappa/\underline{\kappa})^{3/2} + \sqrt{s\mathrm{E}[\rho^2]/\underline{\kappa}}\ \sqrt{\bar\kappa/\underline{\kappa}}(1 + 2\sqrt{\bar\kappa/\underline{\kappa}})\} + \\ &\quad + \|x_i\|_\infty(\sqrt{s^2/n}\sqrt{\bar\kappa/\underline{\kappa}}\ a/[a-1] + 2\sqrt{s\mathrm{E}[\rho^2]/\underline{\kappa}}) + |\rho_i| \\ &\leqslant \|x_i\|_\infty(\bar\kappa/\underline{\kappa})^{3/2}\{\tfrac{2a-1}{a-1}\sqrt{s^2/n} + 5\sqrt{s\mathrm{E}[\rho^2]/\underline{\kappa}}\} + |\rho_i|. \end{aligned}$$

□

F.2. **Verification for Example 2.** Let $\mathbf{P}$ be the collection of all regression models P that obey the conditions set forth above for all n for the given constants $(\underline{\kappa}, \bar\kappa, a, A, B, \chi)$ and sequences p_n and $\bar\delta_n$. Below we provide explicit bounds for κ', κ'', c, C, δ_n and Δ_n that appear in Conditions ASTE, SE and SM that depend only on $(\underline{\kappa}, \bar\kappa, a, A, B, \chi)$, p, $\bar\delta_n$ and n which in turn establish these conditions for any $\mathrm{P} \in \mathbf{P}$. In what follows we exploit Gaussianity of w_i

and use that $(\mathrm{E}[|\eta' w_i|^k])^{1/k} \leqslant G_k(\mathrm{E}[|\eta' w_i|^2])^{1/2}$ for any vector η, $\|\eta\| < \infty$, where the constant G_k depends on k only.

Conditions ASTE(i) is assumed. Condition ASTE(ii) holds with $\|\alpha_0\| \leqslant B =: C_1^{ASTE}$. Because $\theta_m, \theta_g \in S_A^a(p)$, Condition ASTE(iii) holds with

$$s = A^{1/a} n^{1/2a}, \quad r_{mi} = m(z_i) - \sum_{j=1}^{p} z_{ij}\beta_{m0j}, \quad \text{and} \quad r_{gi} = g(z_i) - \sum_{j=1}^{p} z_{ij}\beta_{g0j}$$

where $\|\beta_{m0}\|_0 \leqslant s$ and $\|\beta_{g0}\|_0 \leqslant s$. Indeed, we have

$$\mathrm{E}[r_{mi}^2] \leqslant \mathrm{E}\left[\left(\sum_{j\geqslant s+1} \theta_{m(j)} z_{i(j)}\right)^2\right] \leqslant \bar\kappa \sum_{j\geqslant s+1} \theta_{m(j)}^2 \leqslant \bar\kappa A^2 s^{-2a+1}/[2a-1] \leqslant \bar\kappa s/n$$

where the first inequality follows by the definition of β_{m0} in (5.26), the second inequality follows from $\theta_m \in S_A^a(p)$, and the last inequality because $s = A^{1/a}n^{1/2a}$. Similarly we have $\mathrm{E}[r_{gi}^2] \leqslant \mathrm{E}[(\sum_{j\geqslant s+1}\theta_{g(j)}z_{i(j)})^2] \leqslant \bar\kappa A^2 s^{-2a+1}/[2a-1] \leqslant \bar\kappa s/n$. Thus let $C_2^{ASTE} := \sqrt{\bar f}$.

Condition ASTE(iv) holds with $\delta_{1n}^{ASTE} := A^{2/a} n^{1/a-1}\log^2(p\vee n) \to 0$ since $s = A^{1/a}n^{1/2a}$, A is fixed, and the assumed condition $n^{(1-a)/a}\log^2(p\vee n)\log^2 n \leqslant \bar\delta_n \to 0$.

The moment restrictions in Condition ASTE(v) are satisfied by the Gaussianity. Indeed, we have for $q = 4/\chi$ (where $\chi < 1$ by assumption)

$$\begin{aligned}
\mathrm{E}[|\tilde\zeta_i|^q] &\leqslant 2^{q-1}\mathrm{E}[|\zeta_i^q|] + 2^{q-1}\mathrm{E}[|r_{gi}^q|] \leqslant 2^{q-1}G_q^q(\mathrm{E}[\zeta_i^2]^{q/2} + \mathrm{E}[r_{gi}^2]^{q/2}) \\
&\leqslant 2^{q-1}G_q^q\{\bar\kappa^{q/2} + \bar\kappa^{q/2}(s/n)^{q/2}\} \\
&\leqslant 2^q G_q^q \bar\kappa^{q/2} =: C_3^{ASTE}
\end{aligned}$$

for $s \leqslant n$, i.e., $n \geqslant A^{2/[2a-1]}$. Similarly, $\mathrm{E}[|\tilde v_i|^q] \leqslant C_3^{ASTE}$. Moreover,

$$\begin{aligned}
|\mathrm{E}[\tilde\zeta_i^2\tilde v_i^2] - \mathrm{E}[\zeta_i^2 v_i^2]| &\leqslant \mathrm{E}[\zeta_i^2 r_{mi}^2] + \mathrm{E}[r_{gi}^2 v_i^2] + \mathrm{E}[r_{mi}^2 r_{gi}^2] \\
&\leqslant \sqrt{\mathrm{E}[\zeta_i^4]\mathrm{E}[r_{mi}^4]} + \sqrt{\mathrm{E}[r_{gi}^4]\mathrm{E}[v_i^4]} + \sqrt{\mathrm{E}[r_{mi}^4]\mathrm{E}[r_{gi}^4]} \\
&\leqslant G_4^2\bar\kappa\mathrm{E}[r_{mi}^2] + G_4^2\bar\kappa\mathrm{E}[r_{gi}^2] + G_4^2\mathrm{E}[r_{mi}^2]\mathrm{E}[r_{gi}^2] \\
&\leqslant G_4^2\bar\kappa^2\{2 + \bar\kappa s/n\}s/n =: \delta_{2n}^{ASTE} \to 0.
\end{aligned}$$

Next note that by Gaussian tail bounds and $\lambda_{\max}(\mathrm{E}[w_i w_i']) \leqslant \bar\kappa$ we have

$$\max_{i\leqslant n}\|x_i\|_\infty \leqslant \sqrt{2\bar\kappa\log(pn)} \text{ with probability at least } 1 - \Delta_{1n}^{ASTE} \tag{F.50}$$

where $\Delta_{1n}^{ASTE} = 1/\sqrt{2\bar\kappa\log(pn)}$. The last requirement in Condition ASTE(v) holds with $q = 4/\chi$

$$\max_{i\leqslant n}\|x_i\|_\infty^2 s n^{-1/2+2/q} \leqslant 2\bar\kappa\log(pn)A^{1/a}n^{\frac{1}{2a}-\frac{1}{2}+\chi/2} =: \delta_{3n}^{ASTE}$$

with probability $1 - \Delta_{1n}^{ASTE}$. By the assumption on a, p, χ, and n, $\delta_{3n}^{ASTE} \to 0$.

To verify Condition SE with $\ell_n = \log n$ note that the minimal and maximal eigenvalues of $\mathrm{E}[x_i x_i']$ are bounded away from zero and from above uniformly in n. By the sub-Gaussianity of the regressors, the result follows by Theorem 3.2 in Rudelson and Zhou (2011) (restated in Lemma 9 in Appendix G) with $\tau = 1/2$, $m = s\log n$, $\alpha = \sqrt{8/3}$. Indeed, for

$$n \geqslant N_n := 80(\alpha^4/\tau^2)(s\log n)\ \log(12ep/[\tau s\log n])$$

we have

$$\kappa' := \underline{\kappa}/4 \leqslant \phi_{\min}(s\log n)[\mathbb{E}_n[x_i x_i']] \leqslant \phi_{\max}(s\log n)[\mathbb{E}_n[x_i x_i']] \leqslant 3\bar{\kappa} =: \kappa''$$

with probability $1-\Delta_{1n}^{SE}$, where $\Delta_{1n}^{SE} = 2\exp(-\tau^2 n/80\alpha^4)$. Note that under ASTE(iv) we have $\Delta_{1n}^{SE} \to 0$ and $n_0^{SE} := \max\{n : n \leqslant N_n\} \leqslant \max\{(12e/\tau)^{2a}A^{-2},\ 80^2(\alpha^8/\tau^4)A^{2/a}, n^*\}$ where n^* is the smallest n such that $\bar{\delta}_n < 1$.

The second and third requirements in Conditions SM(i) holds by the Gaussianity of w_i, $\mathrm{E}[\zeta_i \mid x_i, v_i] = 0$, $\mathrm{E}[v_i \mid x_i] = 0$, and the assumption on that the minimal and maximum eigenvalues of the covariance matrix (operator) $\mathrm{E}[w_i w_i']$ are bounded below and above by positive absolute constants.

The first requirement in Condition SM(i) and Condition SM(ii) also hold by Gaussianity. Indeed, we have for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$, $\tilde{y}_i = d_i$ and $\tilde{y}_i = y_i$

$$\begin{aligned}
\mathrm{E}[|v_i^q|] + \mathrm{E}[|\zeta_i^q|] &\leqslant 2^{q-1}G_q^q\{(\mathrm{E}[v_i^2])^{q/2} + (\mathrm{E}[\zeta_i^2])^{q/2}\} \leqslant 2^q G_q^q \bar{\kappa}^{q/2} =: A_1 \\
\mathrm{E}[|d_i^q|] &\leqslant 2^{q-1}\mathrm{E}[|\theta_m' z|^q] + 2^{q-1}\mathrm{E}[|v_i^q|] \leqslant 2^{q-1}G_q^q(\mathrm{E}[|\theta_m' z|^2])^{q/2} + 2^{q-1}G_q^q(\mathrm{E}[v_i^2])^{q/2} \\
&\leqslant 2^{q-1}G_q^q\|\theta_m\|^q\bar{\kappa}^{q/2} + 2^{q-1}G_q^q\bar{\kappa}^{q/2} \leqslant 2^q G_q^q\bar{\kappa}^{q/2}(1+(2A)^q) =: A_2 \\
\mathrm{E}[d_i^2] &\leqslant 2\mathrm{E}[|\theta_m' z_i|^2] + 2\mathrm{E}[v_i^2] \leqslant 2\bar{\kappa}\|\theta_m\|^2 + 2\bar{\kappa} \leqslant 2\bar{\kappa}(4A^2+1) =: A_2' \\
\mathrm{E}[y_i^2] &\leqslant 3|\alpha_0|^2\mathrm{E}[d_i^2] + 3\mathrm{E}[|\theta_m' z|^2] + 3\mathrm{E}[\zeta_i^2] \leqslant 3B^2 A_2' + 3A_2' + 3\bar{\kappa} =: A_3 \\
\max_{1\leqslant j\leqslant p}\mathrm{E}[x_{ij}^2\tilde{y}_i^2] &\leqslant \max_{1\leqslant j\leqslant p}(\mathrm{E}[x_{ij}^4])^{1/2}(\mathrm{E}[\tilde{y}_i^4])^{1/2} \leqslant G_4^4\max_{1\leqslant j\leqslant p}\mathrm{E}[x_{ij}^2]\mathrm{E}[\tilde{y}_i^2] \\
&\leqslant G_4^4\bar{\kappa}(A_2' \vee A_3) =: A_4 \\
\max_{1\leqslant j\leqslant p}\mathrm{E}[|x_{ij}\epsilon_i|^3] &\leqslant \max_{1\leqslant j\leqslant p}(\mathrm{E}[x_{ij}^6])^{1/2}(\mathrm{E}[\epsilon_i^6])^{1/2} \leqslant G_6^6\max_{1\leqslant j\leqslant p}(\mathrm{E}[x_{ij}^2])^{3/2}(\mathrm{E}[\epsilon_i^2])^{3/2} \\
&\leqslant G_6^6\bar{\kappa}^3 =: A_5 \\
\max_{1\leqslant j\leqslant p} 1/\mathrm{E}[x_{ij}^2] &\leqslant 1/\lambda_{\min}(\mathrm{E}[w_i w_i']) \leqslant 1/\underline{\kappa} =: A_6
\end{aligned}$$

because $\|\theta_m\| \leqslant 2A$ and $\|\theta_g\| \leqslant 2A$ since $\theta_m, \theta_g \in S_A^a(p)$. Thus the first requirement in Condition SM(i) holds with $C_2^{SM} = A_2$. Condition SM(ii) holds with $C_3^{SM} = A_1 + (A_2' \vee A_3) + A_4 + A_5 + A_6$.

Condition SM(iii) is assumed.

To verify Condition SM(iv) note that for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$, by (F.50), with probability $1 - \Delta_{1n}^{ASTE}$,

$$\begin{aligned} \max_{j\leqslant p} \sqrt{\mathbb{E}_n[x_{ij}^4\epsilon_i^4]} &\leqslant \max_{j\leqslant p} \sqrt[4]{\mathbb{E}_n[x_{ij}^8]}\sqrt[4]{\mathbb{E}_n[\epsilon_i^8]} \\ &\leqslant \sqrt{2\bar\kappa\log(pn)}\ \max_{j\leqslant p} \sqrt[4]{\mathbb{E}_n[x_{ij}^4]}\sqrt[4]{\mathbb{E}_n[\epsilon_i^8]}. \end{aligned} \tag{F.51}$$

By Lemma 3 with $k = 4$ we have with probability $1 - \Delta_{1n}^{SM}$, where $\Delta_{1n}^{SM} = 1/n$

$$\max_{j\leqslant p} \sqrt[4]{\mathbb{E}_n[x_{ij}^4]} \leqslant \sqrt{\bar\kappa}\sqrt{2}e + \sqrt{\bar\kappa}n^{-1/4}\sqrt{2\log(2pn)} \leqslant 5\sqrt{\bar\kappa} \tag{F.52}$$

for $n \geqslant n_{01}^{SM} = 4\log^2(2pn)$. Also, Lemma 3 with $k = 8$ and $p = 1$ we have with probability $1 - \Delta_{1n}^{SM}$ that

$$\sqrt[4]{\mathbb{E}_n[\epsilon_i^8]} \leqslant 2\bar\kappa 2e^2 + 2\bar\kappa n^{-1/4}2\log n \leqslant 5e^2\bar\kappa \tag{F.53}$$

for $n \geqslant n_{02}^{SM} = 16\log^4 n$. Moreover, we have

$$\max_{1\leqslant j\leqslant p} \sqrt{\mathrm{E}[x_{ij}^4\epsilon_i^4]} \leqslant \max_{1\leqslant j\leqslant p} \sqrt[4]{\mathrm{E}[x_{ij}^8]}\sqrt[4]{\mathrm{E}[\epsilon_i^8]} \leqslant G_8^4\bar\kappa^2.$$

Applying Lemma 6, for $\tau = 2\Delta_{1n}^{ASTE} + \Delta_{2n}^{SM}$, with probability $1 - 8\tau$ we have

$$\max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2\epsilon_i^2]| \leqslant 4\sqrt{\frac{2\log(2p/\tau)}{n}}\sqrt{Q(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[x_{ij}^4\epsilon_i^4], 1-\tau)} \vee 2\sqrt{2}G_8^4\bar\kappa^2/\sqrt{n}$$

where by (F.51), (F.52) and (F.53) we have

$$Q(\max_{1\leqslant j\leqslant p} \sqrt{\mathbb{E}_n[x_{ij}^4\epsilon_i^4]}, 1-\tau) \quad \leqslant \bar\kappa^2\sqrt{2\log(pn)}25e^2.$$

So we let $\delta_{1n}^{SM} = 200e^2\bar\kappa^2\sqrt{\frac{\log(2p/\tau)}{n}}\sqrt{\log(pn)} \vee 2\sqrt{2}\frac{G_8^4\bar\kappa^2}{\sqrt{n}} \to 0$ under the condition that $\log^2(p\vee n)/n \leqslant \bar\delta_n$.

Similarly for $\tilde y_i = d_i$ and $\tilde y_i = y_i$, by Lemma 3, we have with probability $1 - \Delta_{1n}^{SM}$, for $n \geqslant n_{02}^{SM}$ we have

$$\mathbb{E}_n[\tilde y_i^8] \leqslant (5e^2\mathrm{E}[\tilde y_i^2])^4 \leqslant (5e^2[A_2' \vee A_3])^4. \tag{F.54}$$

Moreover, $\sqrt[4]{\mathrm{E}[\tilde y_i^8]} \leqslant G_8^2\mathrm{E}[\tilde y_i^2] \leqslant G_8^2[A_2' \vee A_3]$. Therefore by Lemma 6, for $\tau = 2\Delta_{1n}^{ASTE} + \Delta_{2n}^{SM}$, with probability $1 - 8\tau$ we have

$$\max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2\tilde y_i^2]| \leqslant 4\sqrt{\frac{2\log(2p/\tau)}{n}}\sqrt{2\log(pn)}5\bar\kappa(5e^2[A_2'\vee A_3]) \vee 2\sqrt{2}\frac{G_8^4\bar\kappa[A_2'\vee A_3]}{\sqrt{n}} =: \delta_{2n}^{SM}$$

where $\delta_{2n}^{SM} \to 0$ under the condition $\log^2(p\vee n)/n \leqslant \bar\delta_n \to 0$.

We have that the last term in Condition SM(iv) satisfies with probability $1 - \Delta_{1n}^{ASTE}$

$$\max \|x_i\|_\infty^2 \frac{s\log(p\vee n)}{n} \leqslant 2\bar\kappa\log(pn)A^{1/a}n^{-1+1/2a}\log(p\vee n) =: \delta_{3n}^{SM}.$$

Under ASTE(iv) we have $\delta_{3n}^{SM} \to 0$.

Finally, we set $n_0 = \max\{n_{01}^{SE}, n_{01}^{SM}, n_{02}^{SM}\}$, $C = \max\{C_1^{ASTE}, C_2^{ASTE}, 2C_3^{ASTE}, C_1^{SM}, C_2^{SM}\}$, $\delta_n = \max\{\delta_{1n}^{ASTE}, \delta_{2n}^{ASTE}, \delta_{1n}^{SM} + \delta_{2n}^{SM} + \delta_{3n}^{SM}\} \to 0$, and $\Delta_n = \max\{33\Delta_{1n}^{ASTE} + 16\Delta_{1n}^{SM}, \Delta_{1n}^{SE}\} \to 0$.

□

Lemma 3. *Let $f_{ij} \sim N(0, \sigma_j^2)$, $\sigma_j \leqslant \sigma$, independent across $i = 1, \ldots, n$, where $j = 1, \ldots, p$. Then for any $k \geqslant 2$ and $\gamma \in (0,1)$ we have*

$$P\left(\max_{1\leqslant j\leqslant p} \{\mathbb{E}_n[|f_{ij}^k|]\}^{1/k} \geqslant \sigma\sqrt{2}e + \sigma n^{-1/k}\sqrt{2\log(2p/\gamma)} \right) \leqslant \gamma.$$

Proof. Note that $P(\mathbb{E}_n[|f_{ij}^k|] > M) = P(\|f_{\cdot j}\|_k^k > Mn) = P(\|f_{\cdot j}\|_k > (Mn)^{1/k})$.

Since $|\|f\|_k - \|g\|_k| \leqslant \|f - g\|_k \leqslant \|f - g\|$, we have that $\|\cdot\|_k$ is 1-Lipschitz for $k \geqslant 2$. Moreover,

$$\begin{aligned} \mathrm{E}[\|f_{\cdot j}\|_k] \leqslant (\mathrm{E}[\|f_{\cdot j}\|_k^k])^{1/k} &= (\sum_{i=1}^n \mathrm{E}[|f_{ij}^k|])^{1/k} = n^{1/k}(\mathrm{E}[|f_{1j}^k|])^{1/k} \\ &= n^{1/k}\{\sigma_j^k 2^{k/2}\Gamma((k+1)/2)\}^{1/k} \leqslant n^{1/k}\sigma\sqrt{2}e \end{aligned}$$

By Ledoux and Talagrand (1991), page 21 equation (1.6), we have

$$P(\|f_{\cdot j}\|_k > (Mn)^{1/k} - \mathrm{E}[\|f_{\cdot j}\|_k]) \leqslant 2\exp(-\{(Mn)^{1/k} - \mathrm{E}[\|f_{\cdot j}\|_k]\}^2/2\sigma_j^2).$$

Setting $M := \{\sigma\sqrt{2}e + \sigma n^{-1/k}\sqrt{2\log(2p/\gamma)}\}^k$, so that $(Mn)^{1/k} = n^{1/k}\sigma\sqrt{2}e + \sigma\sqrt{2\log(2p/\gamma)}$ we have by the union bound and $\sigma \geqslant \sigma_j$

$$P(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[|f_{ij}^k|] \geqslant M) \leqslant p \max_{1\leqslant j\leqslant p} P(\mathbb{E}_n[|f_{ij}^k|] \geqslant M) \leqslant \gamma.$$

□

F.3. **Verification for Example 3.** Let **P** be the collection of all regression models P that obey the conditions set forth above for all n for the given constants $(\underline{f}, \bar{f}, a, A, b, B, q)$. Below we provide explicit bounds for κ', κ'', c, C, δ_n and Δ_n that appear in Conditions ASTE, SE and SM that depend only on $(\underline{f}, \bar{f}, a, A, b, B, q)$, p, $\bar{\delta}_n$, and n which in turn establish these conditions for any $\mathrm{P} \in \mathbf{P}$.

Conditions ASTE(i) is assumed. Condition ASTE(ii) holds with $\|\alpha_0\| \leqslant B =: C_1^{ASTE}$. Because $\theta_m, \theta_g \in S_A^a(p)$, Condition ASTE(iii) holds with

$$s = A^{1/a} n^{\frac{1}{2a}}, \quad r_{mi} = m(z_i) - \sum_{j=1}^p \beta_{m0j} P_j(z_i) \;\text{ and }\; r_{gi} = g(z_i) - \sum_{j=1}^p \beta_{g0j} P_j(z_i)$$

where $\|\beta_{m0}\|_0 \leqslant s$ and $\|\beta_{g0}\|_0 \leqslant s$. Indeed, we have

$$\mathrm{E}[r_{mi}^2] \leqslant \mathrm{E}\left[\left(\sum_{j\geqslant s+1}\theta_{m(j)}P_{(j)}(z_i)\right)^2\right] \leqslant \bar{f}\sum_{j\geqslant s+1}\theta_{m(j)}^2 \leqslant \bar{f}A^2 s^{-2a+1}/[2a-1] = \bar{f}s/n$$

where the first inequality follows by the definition of β_{m0} in (5.26), the second inequality follows from the upper bound on the density and orthogonality of the basis, the third inequality follows from $\theta_m \in S_A^a(p)$, and the last inequality because $s = A^{1/a}n^{1/2a}$. Similarly we have $\mathrm{E}[r_{gi}^2] \leqslant \mathrm{E}[(\sum_{j\geqslant s+1}\theta_{g(j)}z_{i(j)})^2] \leqslant \bar{f}A^2s^{-2a+1}/[2a-1] = \bar{f}s/n$. Let $C_2^{ASTE} = \sqrt{\bar{f}}$.

Condition ASTE(iv) holds with $\delta_{1n}^{ASTE} := A^{2/a}n^{1/a-1}\log^2(p\vee n) \to 0$ since $s = A^{1/a}n^{1/2a}$, A is fixed, and the assumed condition $n^{(1-a)/a}\log^2(p\vee n) \leqslant \bar{\delta}_n \to 0$.

To show that Condition ASTE(v) first recall that $\max_{i\leqslant n}\|x_i\|_\infty \leqslant B$. Because $\underline{f} \leqslant \lambda_{\min}(\mathrm{E}[x_ix_i']) \leqslant \lambda_{\max}(\mathrm{E}[x_ix_i']) \leqslant \bar{f}$, by the assumption on the density and orthonormal basis, by Lemma 2 with $\rho = 0$ we have

$$\max_{1\leqslant i\leqslant n}|r_{mi}|\vee|r_{gi}| \leqslant \max_{1\leqslant i\leqslant n}\|x_i\|_\infty(\bar{f}/\underline{f})^{3/2}\frac{2a-1}{a-1}\sqrt{s^2/n} \leqslant B(\bar{f}/\underline{f})^{3/2}\frac{2a-1}{a-1}\sqrt{s^2/n} =: \delta_{2n}^{ASTE}$$

where $\delta_{2n}^{ASTE} \to 0$ under $s = A^{1/a}n^{1/2a}$ and $a > 1$.

Next we establish the moment restrictions in Condition ASTE(v) First, we have

$$\begin{aligned}\mathrm{E}[|\tilde{\zeta}_i|^q] &\leqslant 2^{q-1}\mathrm{E}[|\zeta_i^q|] + 2^{q-1}\mathrm{E}[|r_{gi}^q|] \leqslant 2^{q-1}B + 2^{q-1}(\delta_{2n}^{ASTE})^q \\ &\leqslant 2^{q-1}B + 2^{q-1}(\delta_{2n_0}^{ASTE})^q =: C_3^{ASTE}.\end{aligned}$$

Similarly, $\mathrm{E}[|\tilde{v}_i|^q] \leqslant C_3^{ASTE}$. Moreover, since $\delta_{2n}^{ASTE} \to 0$ we have

$$\begin{aligned}|\mathrm{E}[\tilde{\zeta}_i^2\tilde{v}_i^2] - \mathrm{E}[\zeta_i^2v_i^2]| &\leqslant \mathrm{E}[\zeta_i^2r_{mi}^2] + \mathrm{E}[r_{gi}^2v_i^2] + \mathrm{E}[r_{mi}^2r_{gi}^2] \\ &\leqslant \sqrt{\mathrm{E}[\zeta_i^4]\mathrm{E}[r_{mi}^4]} + \sqrt{\mathrm{E}[r_{gi}^4]\mathrm{E}[v_i^4]} + \sqrt{\mathrm{E}[r_{mi}^4]\mathrm{E}[r_{gi}^4]} \\ &\leqslant 2B^{2/q}(\delta_{2n}^{ASTE})^2 + (\delta_{2n}^{ASTE})^4 =: \delta_{3n}^{ASTE} \to 0.\end{aligned}$$

Finally, the last requirement holds because $(1-a)/a + 4/q < 0$ implies

$$\max_{i\leqslant n}\|x_i\|_\infty^2 sn^{-1/2+2/q} \leqslant B^2A^{1/a}n^{1/2a-1/2+2/q} =: \delta_{4n}^{ASTE} \to 0,$$

since $s = A^{1/a}n^{1/2a}$ and $\max_{i\leqslant n}\|x_i\|_\infty \leqslant B$.

To show Condition SE with $\ell_n = \log n$ note that regressors are uniformly bounded, and minimal and maximal eigenvalues of $\mathrm{E}[x_ix_i']$ are bounded below by $\underline{f}$ and above by $\bar{f}$ uniformly in n. Thus Condition SE follows by Corollary 4 in the supplementary material in Belloni and Chernozhukov (2011) (restated in Lemma 8 in Appendix G) which is based on Rudelson and Vershynin (2008). Let

$$\delta_{1n}^{SE} := 2\bar{C}B\sqrt{s\log n}\log(s\log n)\sqrt{\log(p\vee n)}\sqrt{\log n}/\sqrt{n}$$

and $\Delta_{1n}^{SE} := (2/\underline{f})(\delta_{1n}^{SE})^2 + \delta_{1n}^{SE}(2\bar{f}/\underline{f})$, where $\bar{C}$ is an universal constant. By this result and the Markov inequality, we have with probability $1-\Delta_{1n}^{SE}$

$$\kappa' := \underline{f}/2 \leqslant \phi_{\min}(s\log n)[\mathbb{E}_n[x_ix_i']] \leqslant \phi_{\max}(s\log n)[\mathbb{E}_n[x_ix_i']] \leqslant 2\bar{f} =: \kappa''.$$

We need to show that $\Delta_{1n}^{SE} \to 0$ which follows from $\delta_{1n}^{SE} \to 0$. We have that

$$\delta_{1n}^{SE} \leqslant \frac{2\bar{C}B(1+A)^2\sqrt{n^{1/2a}}\log^2(n)\sqrt{\log(p\vee n)}}{\sqrt{n}} = 2\bar{C}B(1+A)^2\sqrt{\frac{n^{1/2a}\log^4 n}{n^{2/3}}}\sqrt{\frac{\log(p\vee n)}{n^{1/3}}}.$$

By assumption we have $\log^3 p/n \leqslant \bar{\delta}_n \to 0$ and $a > 1$ we have $\delta_{1n}^{SE} \to 0$.

The second and third requirements in Condition SM(i) hold with $C_1^{SM} = B^{2/q}$ and $c_1^{SM} = b$ by assumption. Condition SM(iii) is assumed.

The first requirement in Condition SM(i) and Condition SM(ii) follow by, for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$, $\tilde{y}_i = d_i$ and $\tilde{y}_i = y_i$

$$\begin{aligned}
\mathrm{E}[|v_i^q|] + \mathrm{E}[|\zeta_i^q|] &\leqslant 2B =: A_1\\
\mathrm{E}[|d_i^q|] &\leqslant 2^{q-1}\mathrm{E}[|\theta_m' x_i|^q] + 2^{q-1}\mathrm{E}[|v_i^q|] \leqslant 2^{q-1}\|\theta_m\|_1^q\mathrm{E}[\|x_i\|_\infty^q] + 2^{q-1}B\\
&\leqslant 2^{q-1}(2A)^qB^q + 2^{q-1}B =: A_2\\
\mathrm{E}[d_i^2] &\leqslant 2\bar{f}\|\theta_m\|^2 + 2\mathrm{E}[v_i^2] \leqslant 8\bar{f}A^2 + 2B^{2/q} =: A_2'\\
\mathrm{E}[y_i^2] &\leqslant 3|\alpha_0|^2\mathrm{E}[d_i^2] + 3\|\theta_g\|_1^2\mathrm{E}[\|x_i\|_\infty^2] + 3\mathrm{E}[\zeta_i^2]\\
&\leqslant 3B^2A_2' + 12A^2B^2 + 3B^{2/q} =: A_3\\
\max_{1\leqslant j\leqslant p}\mathrm{E}[x_{ij}^2\tilde{y}_i^2] &\leqslant B^2\mathrm{E}[\tilde{y}_i^2] \leqslant B^2(A_2'\vee A_3) =: A_4\\
\max_{1\leqslant j\leqslant p}\mathrm{E}[|x_{ij}\epsilon_i|^3] &\leqslant B^3\mathrm{E}[|\epsilon_i^3|] \leqslant B^3B^{3/q} =: A_5\\
\max_{1\leqslant j\leqslant p} 1/\mathrm{E}[x_{ij}^2] &\leqslant 1/\lambda_{\min}(\mathrm{E}[x_ix_i']) \leqslant 1/\underline{f} =: A_6
\end{aligned}$$

where we used that $\max_{i\leqslant n}\|x_i\|_\infty \leqslant B$, the moment assumptions of the disturbances, $\|\theta_m\| \leqslant \|\theta_m\|_1 \leqslant 2A$, $\|\theta_g\|_1 \leqslant 2A$ since $\theta_m, \theta_g \in S_A^a(p)$ for $a > 1$. Thus the first requirement in Condition SM(i) holds with $C_2^{SM} = A_2$. Condition SM(ii) holds with $C_3^{SM} := A_1 + (A_2' \vee A_3) + A_4 + A_5 + A_6$.

To verify Condition SM(iv) note that for $\epsilon_i = v_i$ and $\epsilon_i = \zeta_i$ we have by Lemma 6 with probability $1-8\tau$, where $\tau = 1/\log n$,

$$\begin{aligned}
\max_{1\leqslant j\leqslant p}|(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2\epsilon_i^2]| &\leqslant 4\sqrt{\tfrac{2\log(2p/\tau)}{n}}Q(\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[x_{ij}^4\epsilon_i^4]}, 1-\tau) \vee 2\max_{1\leqslant j\leqslant p}\sqrt{2\mathrm{E}[x_{ij}^4\epsilon_i^4]}\\
&\leqslant 4\sqrt{\tfrac{2\log(2p/\tau)}{n}}B^2Q(\sqrt{\mathbb{E}_n[\epsilon_i^4]}, 1-\tau) \vee 2B^2\sqrt{2\mathrm{E}[\epsilon_i^4]}\\
&\leqslant 4\sqrt{\tfrac{2\log(2p\log n)}{n}}B^2B^{2/q}\log n =: \delta_{1n}^{SM}
\end{aligned}$$

where we used $\mathrm{E}[\epsilon_i^4] \leqslant B^{4/q}$ and the Markov inequality. By the definition of τ and the assumed rate $\log^3(p\vee n)/n \leqslant \bar{\delta}_n \to 0$, we have $\delta_{1n}^{SM} \to 0$.

Similarly, we have for $\tilde{y}_i = d_i$ and $\tilde{y}_i = y_i$, with probability $1 - 8\tau$

$$\begin{aligned}\max_{1\leqslant j\leqslant p}|(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2\tilde{y}_i^2]| &\leqslant 4\sqrt{\tfrac{2\log(2p/\tau)}{n}}Q(\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[x_{ij}^4\tilde{y}_i^4]}, 1-\tau) \vee 2\max_{1\leqslant j\leqslant p}\sqrt{2\mathrm{E}[x_{ij}^4\tilde{y}_i^4]}\\ &\leqslant 4\sqrt{\tfrac{2\log(2p/\tau)}{n}}B^2Q(\sqrt{\mathbb{E}_n[\tilde{y}_i^4]}, 1-\tau) \vee 2B^2\sqrt{2\mathrm{E}[\tilde{y}_i^4]}\\ &\leqslant 4\sqrt{\tfrac{2\log(2p\log n)}{n}}B^2A_7\log n =: \delta_{2n}^{SM}\end{aligned}$$

where we used the Markov inequality and

$$\begin{aligned}\mathrm{E}[\tilde{y}_i^4] &\leqslant \mathrm{E}[d_i^4] + 3^3|\alpha_0|^4\mathrm{E}[d_i^4] + 3^3\|\theta_g\|_1^4\mathrm{E}[\|x_i\|_\infty^4] + 3^3\mathrm{E}[\zeta_i^4]\\ &\leqslant A_2^{4/q} + 3^3B^4A_2^{4/q} + 3^3(2A)^4B^4 + 3^3B^{4/q} =: A_7.\end{aligned}$$

By the definition of τ and the assumed rate $\log^3(p\vee n)/n \leqslant \bar{\delta}_n \to 0$, we have $\delta_{2n}^{SM} \to 0$.

The last term in the requirement of Condition SM(iv), because $\max_{i\leqslant n}\|x_i\|_\infty \leqslant B$ and Condition ASTE(iv) holds, is bounded by $\delta_{3n}^{SM} := B^2A^{1/a}n^{1/2a}\log(p\vee n)/n \to 0$.

Finally, we set $c = c_1^{SM}$, $C = \max\{C_1^{ASTE}, C_2^{ASTE}, 2C_3^{ASTE}, C_1^{SM}, C_2^{SM}, C_3^{SM}\}$, $\delta_n = \max\{\bar{\delta}_n, \delta_{1n}^{ASTE}, \delta_{2n}^{ASTE}, \delta_{3n}^{ASTE}, \delta_{4n}^{ASTE}, \delta_{1n}^{SM} + \delta_{2n}^{SM} + \delta_{3n}^{SM}\} \to 0$, $\Delta_n = \max\{16/\log n, \Delta_{1n}^{SE}\} \to 0$. □

F.4. **Verification for Example 4.** The verification of Conditions ASTE, SE, and SM follows as in Example 3. Here we point out the few needed adjustments. To show Condition ASTE(iii), the sparsity condition, follows from $\theta_m, \theta_g \in S_A^a(p)$. Indeed, we have

$$r_{mi} = m(z_i) - \sum_{j=1}^p \beta_{m0j}P_{j,p}(z_i) \ \text{ and } \ r_{gi} = g(z_i) - \sum_{j=1}^p \beta_{g0j}P_{j,p}(z_i)$$

where $\|\beta_{m0}\|_0 \leqslant s$ and $\|\beta_{g0}\|_0 \leqslant s$, so that

$$\begin{aligned}\sqrt{\mathrm{E}[r_{mi}^2]} &\leqslant \sqrt{\mathrm{E}[(\textstyle\sum_{j=s+1}^p \theta_{g(j)}P_{(j),p}(z_i))^2]} + \sqrt{\mathrm{E}[\rho_{m,p}^2(z_i)]}\\ &\leqslant \sqrt{(\bar{f}/[2a-1])s/n} + Ap^{-a+1/2} \leqslant \sqrt{\bar{f}s/n} + (\bar{\delta}_n)^{a-1/2}\sqrt{s/n},\end{aligned}$$

since $a > 1$, $s = A^{1/a}n^{1/2a} \leqslant p\bar{\delta}_n$, where $\bar{\delta}_n \to 0$. Similarly $\sqrt{\mathrm{E}[r_{gi}^2]} \leqslant \sqrt{\bar{f}s/n} + (\bar{\delta}_n)^{a-1/2}\sqrt{s/n}$.

Also, because $\max_{i\leqslant n}\|x_i\|_\infty \leqslant B$ and $\underline{f} \leqslant \lambda_{\min}(\mathrm{E}[x_ix_i']) \leqslant \lambda_{\max}(\mathrm{E}[x_ix_i']) \leqslant \bar{f}$ by the assumption on the density and orthonormal basis, by Lemma 2 with $\rho = \rho_{m,p}$ where $\sqrt{\mathrm{E}[\rho^2]} \leqslant Ap^{-a+1/2}$ and $|\rho| \leqslant Ap^{-a+1}$, we have

$$|r_{mi}| \leqslant B(\bar{f}/\underline{f})^{3/2}\left\{\frac{2a-1}{a-1}\sqrt{s^2/n} + 5\sqrt{s/\underline{\kappa}}Ap^{-a+1/2}\right\} + Ap^{-a+1} =: \delta_{1n}^{ASTE} \to 0$$

since $s^2/n \to 0$, $s \leqslant p\bar{\delta}_n$, $a > 1$ and $s \to \infty$. □

Appendix G. Tools

G.1. **Moderate Deviations for a Maximum of Self-Normalized Averages.** We shall be using the following result, which is based on Theorem 7.4 in (de la Peña, Lai, and Shao, 2009).

Lemma 4 (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 $\ell_n > 0$

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

The proof of this result, given in Belloni, Chen, Chernozhukov, and Hansen (2010), follows from a simple combination of union bounds with the bounds in Theorem 7.4 in de la Peña, Lai, and Shao (2009).

G.2. **Inequalities based on Symmetrization.** Next we proceed to use symmetrization arguments to bound the empirical process. In what follows for a random variable Z let $Q(Z, 1-\tau)$ denote its $(1-\tau)$-quantile.

Lemma 5 (Maximal inequality via symmetrization)**.** *Let $Z_1, \ldots, Z_n$ be arbitrary independent stochastic processes and $\mathcal{F}$ a finite set of measurable functions. For any $\tau \in (0, 1/2)$, and $\delta \in (0,1)$ we have that with probability at least $1-4\tau-4\delta$*

$$\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f(Z_i))| \leqslant \left\{4\sqrt{2\log(2|\mathcal{F}|/\delta)}\; Q\left(\sup_{f\in\mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]}, 1-\tau\right)\right\} \vee 2\max_{f\in\mathcal{F}} Q\left(|\mathbb{G}_n(f(Z_i))|, \frac{1}{2}\right).$$

Proof. Let

$$e_{1n} = \sqrt{2\log(2|\mathcal{F}|/\delta)}\; Q\left(\sup_{f\in\mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]}, 1-\tau\right), \quad e_{2n} = \max_{f\in\mathcal{F}} Q\left(|\mathbb{G}_n(f)|, \frac{1}{2}\right)$$

and the event $\mathcal{E} = \{\sup_{f\in\mathcal{F}} \sqrt{\mathbb{E}_n[f^2(Z_i)]} \leqslant Q\left(\sup_{f\in\mathcal{F}} \sqrt{\mathbb{E}_n[f^2(Z_i)]}, 1-\tau\right)\}$ which satisfies $P(\mathcal{E}) \geqslant 1-\tau$. By the symmetrization Lemma 2.3.7 of van der Vaart and Wellner (1996) (by definition of e_{2n} we have $\beta_n(x) \geqslant 1/2$ in Lemma 2.3.7) we obtain

$$\begin{array}{rl} \mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f(Z_i))| > 4e_{1n} \vee 2e_{2n}\right\} & \leqslant 4\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\mathbb{G}_n(\varepsilon_i f(Z_i))| > e_{1n}\right\} \\ & \leqslant 4\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\mathbb{G}_n(\varepsilon_i f(Z_i))| > e_{1n}|\mathcal{E}\right\} + 4\tau \end{array}$$

where ε_i are independent Rademacher random variables, $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$.

Thus a union bound yields

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f(Z_i))| > 4e_n(\mathcal{F})\right\} \leqslant 4\tau + 4|\mathcal{F}| \sup_{f\in\mathcal{F}} \mathbb{P}\left\{|\mathbb{G}_n(\varepsilon_i f(Z_i))| > e_{1n}|\mathcal{E}\right\}. \tag{G.55}$$

We then condition on the values of $Z_1, \ldots, Z_n$ and $\mathcal{E}$, denoting the conditional probability measure as $\mathbb{P}_\varepsilon$. Conditional on $Z_1, \ldots, Z_n$, by the Hoeffding inequality the symmetrized process $\mathbb{G}_n(\varepsilon_i f(Z_i))$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$ norm, namely, for $f \in \mathcal{F}$, $\mathbb{P}_\varepsilon\{|\mathbb{G}_n(\varepsilon_i f(Z_i))| > x\} \leqslant 2\exp(-x^2/[2\mathbb{E}_n[f^2(Z_i)]])$. Hence, under the event $\mathcal{E}$, we can bound

$$\begin{array}{rcl} \mathbb{P}_\varepsilon\left\{|\mathbb{G}_n(\varepsilon_i f(Z_i))| > e_{1n}|Z_1, \ldots, Z_n, \mathcal{E}\right\} & \leqslant & 2\exp(-e_{1n}^2/[2\mathbb{E}_n[f^2(Z_i)]]) \\ & \leqslant & 2\exp(-\log(2|\mathcal{F}|/\delta)). \end{array}$$

Taking the expectation over $Z_1, \ldots, Z_n$ does not affect the right hand side bound. Plugging in this bound yields the result. □

The following specialization will be convenient.

Lemma 6. *Let $\{(x_i', \epsilon_i)' \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, n\}$ be random vectors that are independent across i. Then with probability at least $1 - 8\tau$*

$$\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[x_{ij}^2\epsilon_i^2] - \bar{\mathrm{E}}[x_{ij}^2\epsilon_i^2]| \leqslant 4\sqrt{\frac{2\log(2p/\tau)}{n}\ Q\left(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[x_{ij}^4\epsilon_i^4], 1-\tau\right)} \vee 2\max_{1\leqslant j\leqslant p} \sqrt{\frac{2\bar{\mathrm{E}}[x_{ij}^4\epsilon_i^4]}{n}}$$

Proof. Let $Z_i = x_i\epsilon_i$, $f_j(Z_i) = x_{ij}^2\epsilon_i^2$, $\mathcal{F} = \{f_1, \ldots, f_p\}$, so that $n^{-1/2}\mathbb{G}_n(f_j(Z_i)) = \mathbb{E}_n[x_{ij}^2\epsilon_i^2] - \bar{\mathrm{E}}[x_{ij}^2\epsilon_i^2]$. Also, let

$$e_{1n} = \sqrt{2\log(2p/\tau_1)}\sqrt{Q\left(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[x_{ij}^4\epsilon_i^4], 1-\tau_2\right)} \ \text{ and } \ e_{2n} = \max_{1\leqslant j\leqslant p} Q(|\mathbb{G}_n(f_j(Z_i))|, 1/2)$$

where we have $e_{2n} \leqslant \max_{1\leqslant j\leqslant p} \sqrt{2\bar{\mathrm{E}}[x_{ij}^4\epsilon_i^4]}$ by Chebyshev.

By Lemma 5 we have

$$P\left(\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[x_{ij}^2\epsilon_i^2] - \bar{\mathrm{E}}[x_{ij}^2\epsilon_i^2]| > \frac{4e_{1n} \vee 2e_{2n}}{\sqrt{n}}\right) \leqslant 4\tau_1 + 4\tau_2.$$

The first result follows by setting $\tau_1 = \tau_2 = \tau < 1/2$. □

G.3. **Moment Inequality.** We shall be using the following result, which is based on Markov inequality and (von Bahr and Esseen, 1965).

Lemma 7 (Vonbahr-Esseen's LLN). *Let $r \in [1,2]$, and independent zero-mean random variables X_i with $\bar{\mathrm{E}}[|X_i|^r] \leqslant C$. Then for any $\ell_n > 0$*

$$Pr\left(\frac{|\sum_{i=1}^n X_i|}{n} > \ell_n n^{-(1-1/r)}\right) \leqslant \frac{2C}{\ell_n^r}.$$

G.4. **Matrices Deviation Bounds.** Based on results in Rudelson and Vershynin (2008), the following lemma for bounded regressors was derived in the supplementary material of Belloni and Chernozhukov (2011).

Lemma 8 (Essentially in Theorem 3.6 of Rudelson and Vershynin (2008)). *Let x_i, $i = 1, \ldots, n$, be i.i.d. random vectors in $\mathbb{R}^p$ with uniformly bounded entries, $\|x_i\|_\infty \leqslant K$ a.s. for all $i = 1, \ldots, n$. Let $\delta_n := 2\left(CK\sqrt{k}\log(k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right)/\sqrt{n}$, where C is the universal constant. Then,*

$$\mathrm{E}\left[\sup_{\|\alpha\|_0 \leqslant k, \|\alpha\|=1} \left|\mathbb{E}_n\left[(\alpha' x_i)^2 - \mathrm{E}[(\alpha' x_i)^2]\right]\right|\right] \leqslant \delta_n^2 + \delta_n \sup_{\|\alpha\|_0 \leqslant k, \|\alpha\|=1} \sqrt{\mathrm{E}[(\alpha' x_i)^2]}.$$

The following result establishes an approximation bound for sub-Gaussian regressors and was developed in Rudelson and Zhou (2011). Recall that a random vector $Z \in \mathbb{R}^p$ is isotropic if $\mathrm{E}[ZZ'] = I$, and it is called ψ_2 with a constant α if for every $w \in \mathbb{R}^p$ we have

$$\|Z'w\|_{\psi_2} := \inf\{t : \mathrm{E}[\exp(\ (Z'w)^2/t^2)] \leqslant 2\} \leqslant \alpha \|w\|_2.$$

Lemma 9 (Essentially in Theorem 3.2 of Rudelson and Zhou (2011)). *Let Ψ_i be i.i.d. isotropic random vectors in $\mathbb{R}^p$ that is ψ_2 with a constant α. Let $x_i = \Sigma^{1/2}\Psi_i$ so that $\Sigma = \mathrm{E}[x_i x_i']$. For $m \leqslant p$ and $\tau \in (0,1)$ assume that*

$$n \geqslant \frac{80 m \alpha^4}{\tau^2} \log\left(\frac{12ep}{m\tau}\right).$$

Then with probability at least $1 - 2\exp(-\tau^2 n / 80\alpha^4)$, for all m-sparse vectors $u \in \mathbb{R}^p$,

$$(1-\tau)\|\Sigma^{1/2}u\|_2 \leqslant \sqrt{\mathbb{E}_n[(x_i'u)^2]} \leqslant (1+\tau)\|\Sigma^{1/2}u\|_2.$$

For example, Lemma 9 covers the case of $x_i \sim N(0, \Sigma)$ by setting $\Psi_i \sim N(0, I)$ which is isotropic and ψ_2 with a constant $\alpha = \sqrt{8/3}$.

References

Abadie, A., and G. W. Imbens (2011): "Bias-corrected matching estimators for average treatment effects," *J. Bus. Econom. Statist.*, 29(1), 1–11.

Andrews, D., and X. Cheng (2011): "Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure," Cowless Foundation Discussion Paper.

Andrews, D., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," Cowless Foundation Discussion Paper.

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

Baraniuk, R., M. Davenport, R. DeVore, and M. Wakin (2008): "A Simple Proof of the Restricted Isometry Property for Random Matrices," *Constructive Approximation*, 28, 253–263.

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

Belloni, A., and V. Chernozhukov (2011): "Least Squares After Model Selection in High-dimensional Sparse Models," *forthcoming Bernoulli*.

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

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

——— (2011b): "Supplementary Material for "Inference on Treatment Effects with High-Dimensional Controls"," *arXiv:1201.0224*.

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

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

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

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

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

Cattaneo, M., M. Jansson, and W. Newey (2010): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Working Paper, http://econ-www.mit.edu/files/6204*.

Chamberlain, G. (1985): "Longitudinal Analysis of Labor Market Data," in *Heterogeneity, Omitted Variable Bias, and Duration Dependence*, ed. by J. J. Heckman, and B. Singer. Cambridge University Press.

Chen, X. (2007): "Large Sample Sieve Estimatin of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5559–5632.

de la Peña, V. H., T. L. Lai, and Q.-M. Shao (2009): *Self-normalized processes*, Probability and its Applications (New York). Springer-Verlag, Berlin, Limit theory and statistical applications.

Donald, S. G., and W. K. Newey (1994): "Series estimation of semilinear models," *J. Multivariate Anal.*, 50(1), 30–40.

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

——— (2008): "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz," *Quarterly Journal of Economics*, 123(1), 425–440.

Fan, J., S. Guo, and N. Hao (2011): "Variance estimation using refitted cross-validation in ultrahigh dimensional regression," *forthcoming Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Fan, J., and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of American Statistical Association*, 96(456), 1348–1360.

Foote, C. L., and C. F. Goetz (2008): "The Impact of Legalized Abortion on Crime: Comment," *Quarterly Journal of Economics*, 123(1), 407–423.

Gautier, E., and A. Tsybakov (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(2), 315–331.

Hansen, B. E. (2005): "Challenges for Econometric Model Selection," *Econometric Theory*, 21, 60–68.

Härdle, W., H. Liang, and J. Gao (2000): *Partially linear models*, Contributions to Statistics. Physica-Verlag, Heidelberg.

Heckman, J., R. LaLonde, and J. Smith (1999): "The economics and econometrics of active labor market programs," *Handbook of labor economics*, 3, 1865–2097.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *Rev. Econom. Stud.*, 65(2), 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36(2), 587613.

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

Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003): "Self-normalized Cramr-type large deviations for independent random variables," *Ann. Probab.*, 31(4), 2167–2215.

Kerkyacharian, G., and D. Picard (1992): "Density estimation in Besov spaces," *Statist. Probab. Lett.*, 13(1), 15–24.

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

Ledoux, M., and M. Talagrand (1991): *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag.

Leeb, H., and B. M. Pötscher (2008): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24(2), 338–376.

MacKinnon, J. G., and H. White (1985): "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

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

Mikusheva, A. (2007): "Uniform inference in autoregressive models," *Econometrica*, 75(5), 1411–1452.

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

Pötscher, B. M. (2009): "Confidence sets based on sparse estimators are necessarily large," *Sankhyā*, 71(1, Ser. A), 1–18.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *J. Amer. Statist. Assoc.*, 90(429), 122–129.

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica*, 56(4), 931–954.

Rudelson, M., and R. Vershynin (2008): "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, 61, 10251045.

Rudelson, M., and S. Zhou (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 der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

von Bahr, B., and C.-G. Esseen (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*.

Table 1. Simulation Results for Selected R^2 Values

Estimation Procedure	First Stage R^2 = .2, Structure R^2 = 0 RMSE	Rej. Rate	First Stage R^2 = .2, Structure R^2 = .8 RMSE	Rej. Rate	First Stage R^2 = .8, Structure R^2 = 0 RMSE	Rej. Rate	First Stage R^2 = .8, Structure R^2 = .8 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).

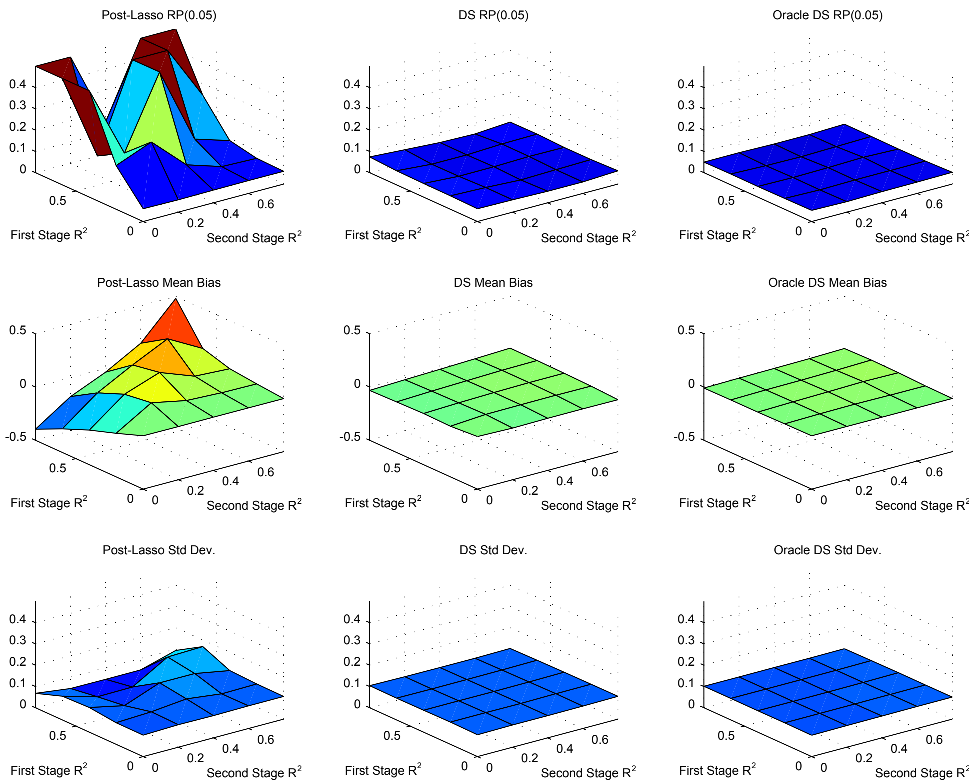

FIGURE 2. This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 1 of the simulation study which has quadratically decaying coefficients and homoscedasticity. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and the infeasible OLS estimator that uses the set of variables that have coefficients larger than 0.1 in either equation (2.6) or (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.

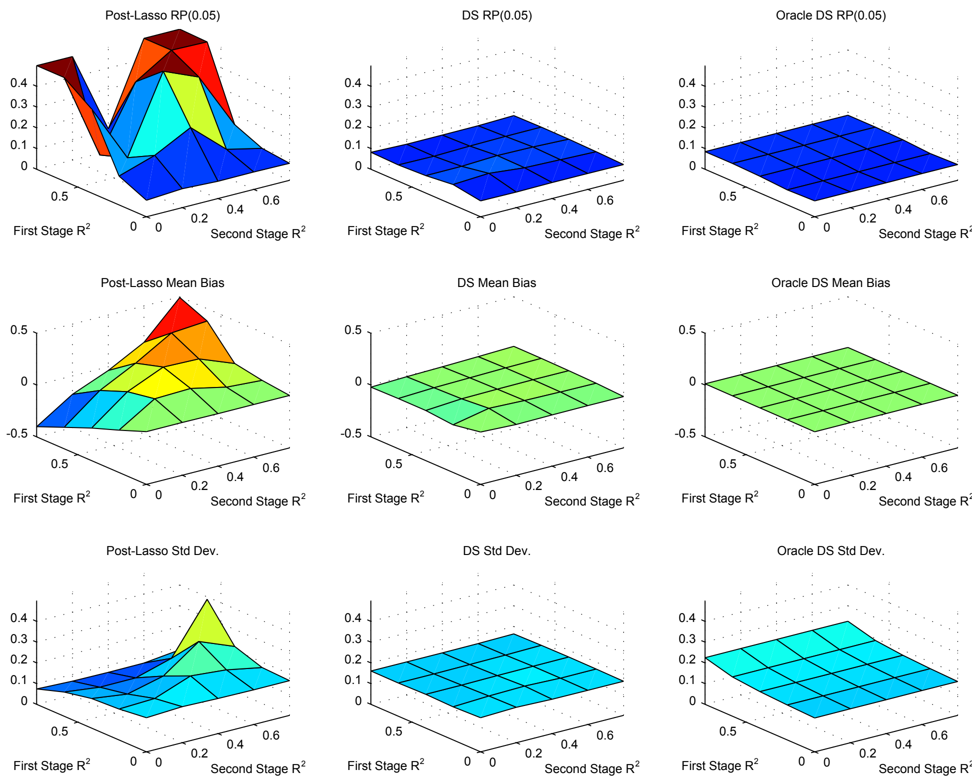

FIGURE 3. This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 2 of the simulation study which has quadratically decaying coefficients and heteroscedasticity. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and the infeasible OLS estimator that uses the set of variables that have coefficients larger than 0.1 in either equation (2.6) or (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.

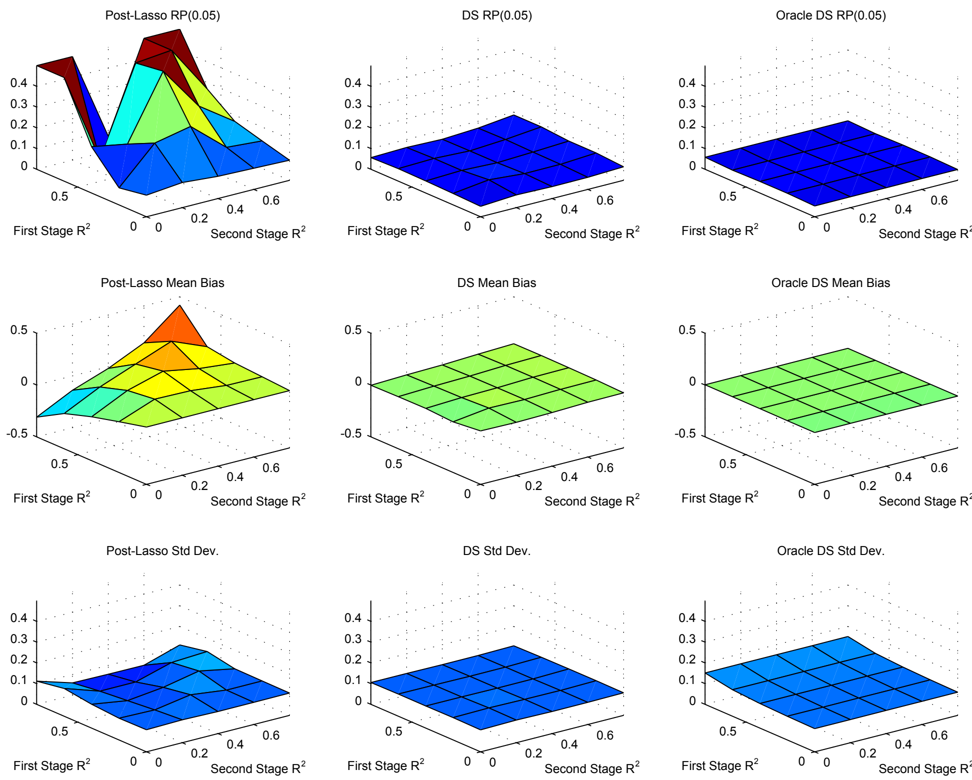

FIGURE 4. This figure presents rejection frequencies for 5% level tests, biases, and standard deviations for estimating the treatment effect from Design 3 of the simulation study which has five quadratically decaying coefficients and 95 Gaussian random coefficients. Results are reported for a one-step Post-Lasso estimator, our proposed double selection procedure, and the infeasible OLS estimator that uses the set of variables that have coefficients larger than 0.1 in either equation (2.6) or (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.

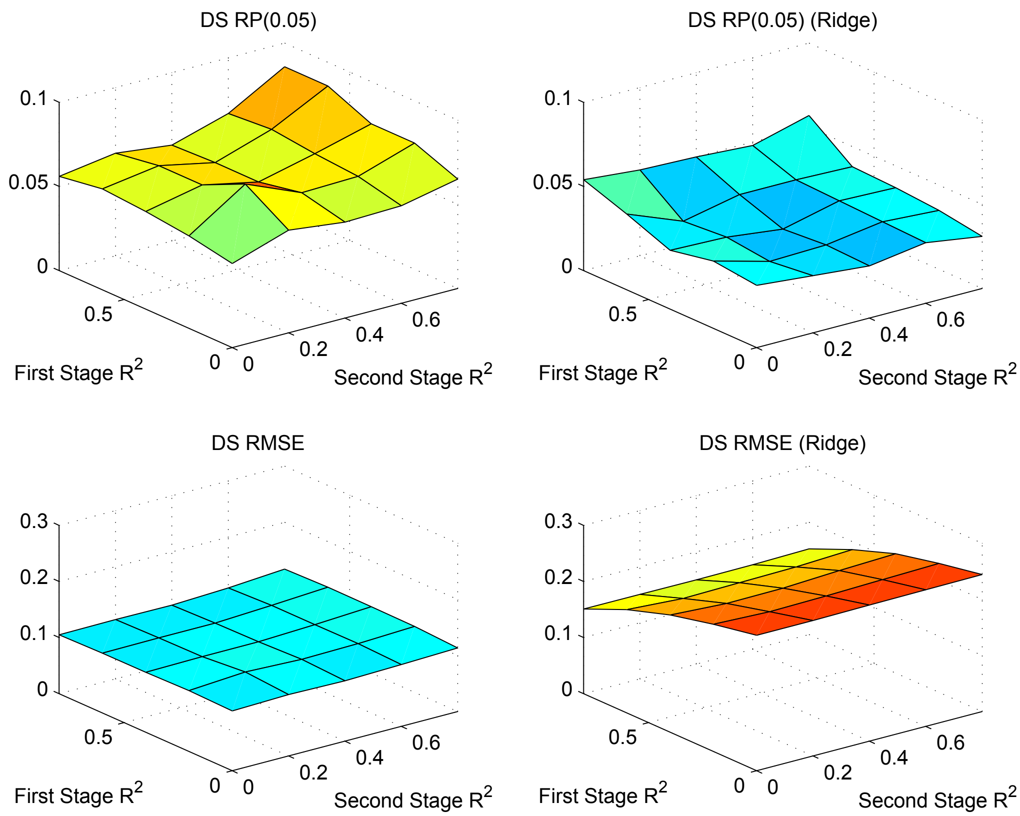

FIGURE 5. This figure presents rejection frequencies for 5% level tests and RMSE's for estimating the treatment effect from Design 3 of the simulation study which has five quadratically decaying coefficients and 95 Gaussian random coefficients. Results in the first column are for the proposed double 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.

Table 2. Estimated Effects of Abortion on Crime Rates (Levels)

	Violent Crime		Property Crime		Murder	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
	A. Donohue and Levitt (2001) Table IV					
DL (2001) Table IV	-0.129	0.024	-0.091	0.018	-0.121	0.047
Fixed Effects	-0.131	0.045	-0.091	0.016	-0.131	0.058
Fixed Effects + State Trends	-0.149	0.185	0.060	0.093	-0.383	0.207
All Controls	0.183	0.447	0.013	0.067	0.855	0.974
Post-Double-Selection	0.133	0.303	-0.053	0.044	-0.692	0.438
Polynomial Trend	0.321	0.349	-0.032	0.060	0.851	0.616
Post-Double-Selection, Polynomial Trend	0.013	0.251	-0.041	0.047	-0.178	0.276
	B. Donohue and Levitt (2008) Table III					
DL (2008) Table III	-0.160	0.088	-0.062	0.030	-0.248	0.100
DL (2008) Specification	-0.158	0.087	-0.057	0.026	-0.249	0.099
Fixed Effects	-0.186	0.063	-0.110	0.046	-0.061	0.078
All Controls	0.516	0.400	0.146	0.127	0.611	0.523
Post-Double-Selection	0.060	0.214	-0.025	0.086	0.460	0.322
Poynomial Trend	0.203	0.296	0.141	0.089	0.199	0.309
Post-Double-Selection, Polynomial Trend	-0.264	0.179	0.090	0.046	-0.088	0.192

Note: The table displays the estimated coefficient on the abortion rate, "Effect," and its estimated standard error. Numbers in the first row of Panel A are taken from Donohue III and Levitt (2001) Table IV, columns (2), (4), and (6). Numbers from the first row of Panel B are taken from Donohue III and Levitt (2008) Table III, column (8). The remaining rows are estimated by OLS of the crime rate on the abortion rate and different sets of controls described in the text and use standard errors clustered at the state-level. In Panel A, the row labeled "All Controls" uses 311 control variables as discussed in the text that include the 68 controls from the original specification of Donohue III and Levitt (2001) Table IV along with 243 variables meant to allow for flexible, smooth trends. The row labeled "Polynomial Trend" in Panel A restricts the set of controls added to allow for flexible trends to include only polynomial terms and uses only 149 total regressors, the 68 from the original specification and 81 added variables. In Panel B, the row labeled "All Controls" uses 713 control variables as discussed in the text that include the 473 controls from the original specification of Donohue III and Levitt (2008) Table III along with 240 variables meant to allow for flexible, smooth trends. The row labeled "Polynomial Trend" in Panel B restricts the set of controls added to allow for flexible trends to include only polynomial terms and uses only 553 total regressors, the 473 from the original specification and 80 added variables. The rows "Post-Double-Selection" report results from regressing the crime rates on the variables from the original Donohue III and Levitt (2001) and Donohue III and Levitt (2008) along with additional variables selected using the technique developed in this paper from among the set of variables considered in the corresponding "All Controls" row. The rows "Post-Double-Selection, Polynomial Trend" report results from regressing the crime rates on the variables from the original Donohue III and Levitt (2001) and Donohue III and Levitt (2008) along with additional variables selected using the technique developed in this paper from among the set of variables considered in the corresponding "Polynomial Trend" row. Further details are provided in the text.