INFERENCE IN ADDITIVELY SEPARABLE MODELS WITH A HIGH DIMENSIONAL COMPONENT

DAMIAN KOZBUR

Abstract. This paper provides inference results for series estimators with a high dimensional component. In conditional expectation models that have an additively separable form, a single component can be estimated with rates customary in nonparametric estimation even when then number of series terms for remaining components is large relative to the sample size. This allows, for example, estimation of a nonlinear response of an outcome variable given a treatment variable of interest while accounting for potentially very many confounders. A key condition which makes inference in this setting possible is sparsity; there is a small (unknown) subset of terms which can replace the entire series without inducing significant bias. This paper considers a model selection procedure for choosing series terms that generalizes the post-double selection procedure given in Belloni, Chernozhukov, Hansen (2013) to the nonparametric setting. In one stage, variables are selected if they are relevant for predicting the treatment. In a second stage, variables are selected if they are relevant in predicting the treatment regressor of interest. Rates of convergence and asymptotic normality are derived for series estimators of a component of a conditional expectation in high dimensional models under sparsity conditions. Simulation results demonstrate that the proposed estimator performs favorably in terms of size of tests and risk properties relative to other estimation strategies.

Key words: nonparametric regression, additively separable, series estimation, high-dimensional models, post-double selection

1. Introduction

Nonparametric estimation of economic or statistical models is useful for applications where functional forms are unavailable. The econometric theory for nonparametric estimation using an approximating series expansion is well-understood under standard regularity conditions; see, for example, Chen (2007), Newey (1997) or Andrews (1991). For many applications, the primary object of interest can be calculated as a conditional expectation function of a response variable y given a regressor x_1 and possible confounders x_2. When x_1 is endogenously determined, or otherwise exhibits dependence with covariates that affect

Date: First version: August 2013, this version December 2, 2013. I gratefully acknowledge helpful comments from Christian Hansen, Matthew Taddy, Azeem Shaikh, Matias Cattaneo, Dan Nguyen, Eric Floyd.

y, estimates of any (nonlinear) partial effect of x_1 on y will be inconsistent if x_2 is ignored. A common method for overcoming this problem is jointly modeling the response of y to x_1 and all other relevant covariates x_2 with the idea that x_1 can be taken as approximately randomly assigned given x_2. This may include, for example, a specifying a partially linear model $E[y|x] = g(x_1) + x_2'\beta$ or a fully nonparametric model $E[y|x] = g(x) = g(x_1, x_2)$. However, the partially linear approach is unreliable or infeasible when the dimension of the potential confounding variables is large. The fully nonparametric model suffers from the curse of dimensionality even for moderately many covariates. This paper gives a formal model selection technique which provides robust inference for the partial effects of x_1 when the dimension of the confounding x_2 variable is prohibitively large for standard methods.

A standard series estimator of $g(x) = E[y|x]$ is obtained with the aid of a dictionary of transformations $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$: a set of K functions of x with the property that a linear combination of the $p_{jK}(x)$ can approximate g to an increasing level of precision that depends on K. K is permitted to depend on n and $p^K(x)$ may include splines, fourier series, orthogonal polynomials or other functions which may be useful for approximating g. The series estimator is simple and implemented with standard least squares regression: given data (y_i, x_i) for $i = 1, ..., n$, a series estimator for g is takes the form: $\widehat{g}(x) = p^K(x)'\widehat{\beta}$ for $P = [p^K(x_1), ..., p^K(x_n)]'$, $Y = (y_1, ..., y_n)'$ and $\widehat{\beta} = (P'P)^{-1}P'Y$. Traditionally, the number of series terms, chosen in a way to simultaneously reduce bias and increase precision, must be small relative to the sample size. Thus the function of interest must be sufficiently smooth or simple. If the dimension of the variable x is high, additional restrictions on the function g are often necessary since approximating arbitrary functions of a high-dimensional variable requires very many terms. A convenient and still flexible restriction on g is that it be additively separable, $g(x) = \sum_{j=1}^{d} g_j(x_j)$ where $(x_1, ..., x_d)$ are components of the vector x. (Stone (1985), Huang, Horowitz and Wei (2010)). This paper focuses on estimation and inference in additively separable models. In particular, when $E[y|x] = g_1(x_1) + g_2(x_2)$, and the target is to learn the function g_1, the structure of an additively separable model allows for inference even when the approximating dictionary for g_2 has a large number of terms.

An alternative to traditional nonparametric estimation using a small number of series terms acting as an approximating model is a sparse high-dimensional approximating model. Sparse approximation generalizes the notion of an approximating series. Sparsity in the context of regression refers to the notion that most parameters are actually zero or very near zero which leaves a small set of nonzero parameters to be estimated. In the context of nonparametric regression, sparsity refers to the notion that a linear combination of a small set of terms approximate the nonparametric function in question. In classical nonparametric

estimiation, the series terms are typically known a priori (though the exact number can be data dependent, for example, choosing K to minimize a cross validation criterion. Sparsity allows the relevant terms to be a priori unknown and estimated from the data.) In this paper, a high-dimensional model refers generally to a model where the number of parameters to be estimated is on the order of, or larger than the sample size. This can mean that there are many distinct variables that enter the conditional expectation function. Alternatively, this can mean that many series terms are required to adequately model a function of a low-dimensional variable. For example, Belloni and Chernozhukov (2011) present methods for nonparametric regression where a small number K terms is selected from a much larger pool of terms using modern variable selection techniques but do not consider inference.

This paper adddresses questions of inference and asymptotic normality for functionals of a component of a nonparametric regression function which is estimated with a series selected by a formal model selection procedure. Modern techniques in high dimensional regression make signal recovery possible in cases where the number of regressors is much higher than the number of observations. By leveraging additively separable structure, inference for nonparametric conditional expectation functions can be performed under much more general approximating series, provided that appropriate sparsity conditions are met. This paper compliments existing literature on nonparametric series estimation by expanding the class of allowable dictionaries when the primary object of interest can still be described by a small number of known series terms.

This paper contributes to a broader program aimed at conducting inference in the context of high-dimensional models. Statistical methods in high dimensions have been well developed for the purpose of prediction (Tibshirani (1996), Hastie, Tibshirani and Friedman (2009) Candes and Tao (2006) Bickel, Ritov, and Tsybakov (2009), Huang, Horowitz, and Wei (2010, Belloni and Chernozkukov (2011), Meinshausen and Yu (2009)). These methods feature regularized estimation which buys stability and reduction in estimate variability at the cost of a modest bias, or estimation which favors parsimony where many parameter values are set identically to zero, or both. More recently, some authors have begun the important task of assigning uncertainties or estimation error to parameter estimates in high dimensional models (Buhlman (2013), Belloni Chernozhukov and Hansen 2013)). Quantifying estimation precision has been shown to be difficult theoretically and in many cases, formal model selection can preclude the validity of standard $\sqrt{n}$ inference (Leeb and Potscher (2008), Potscher (2009)). The paper builds on the methodology found in Belloni, Chernozhukov and Hansen (2013) which give robust statistical inference for the slope parameter of a treatment variable d with high-dimensional confounders z by selecting the

elements of z that are most useful for predicting d in one step, and selecting elements of z most useful for predicting y in a second step. The use of two model selection steps overcomes impossibility results about statistical inference. This paper generalizes the approach from estimating a linear treatment model to estimating a component in nonparametric additively separable models. The main technical contribution lies in providing conditions under which model selection provides inference that is uniformly robust to suitably regular functions.

2. A high dimensional additively separable model

This section provides an intuitive discussion of the additively separable nonparametric model explored in this paper. Consider a conditional expectation function with two distinct components:

$$E[y|x] = g(x) = g_1(x_1) + g_2(x_2)$$

The component functions g_1 and g_2 are restricted to belong to ambient spaces $g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2$ which allow them to be uniquely identified. The function g and therefore, g_1, g_2 will be allowed to depend on n which allows for a many variable setup, however, this dependence will be supressed from the notation. In particular, this allows estimation of models of the form $E[y|x] = g_1(x_1) + x_{2n}'\beta_n$ with $\dim(x_{2n}) = p_n$. This is useful for modeling nonlinear conditional expectation functions with a large list of potential confounders. The formulation will be slightly more general than the additively separable model. It allows, for instance, additive interaction models like those found in Andrews and Whang (1991) so that. For example, the model $E[y|x] = g_1(x_1) + \gamma \cdot x1 \cdot x2 + g_2(x_2)$ where γ is a parameter to be estimated is allowed. Then $\mathcal{G}_1$ consists of functions that depend only on x_1 except for the additional term $\gamma \cdot x_1 \cdot x_2$. Alternatively, x_1 and x_2 can share components, provided that the researcher provides conditions on $\mathcal{G}_1$ and $\mathcal{G}_2$ that allow g_1 and g_2 to be well identified. Because of this, and for the sake of notation, $g_1(x_1)$, $g_2(x_2)$ will simply be written, $g_1(x)$, $g_2(x)$.

The estimation of g proceeds by a series approximation. Suppose there is a dictionary $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))$ which is compatible with the decomposition given above so that p^K can be separated into two parts $p^K = (p_1^{K_1}, p_2^{K_2})$. The approximating dictionaries $p_1^{K_1}(x) = (p_{1;1K_1}(x), ..., p_{1;K_1K_1}(x))'$ and $p_2^{K_2}(x) = (p_{2;1K_2}(x), ..., p_{2;K_2K_2}(x))'$ are for the two components g_1, g_2. In what follows, dependence on K_1 and K_2 is suppressed in the notation so that for example $p_1^{K_1}(x) = p_1(x)$ and $p_2^{K_2}(x) = p_2(x)$.

The two dictionaries differ in nature. The first dictionary, $p_1(x)$ is traditional, and follows standard conditions imposed on series estimators, for example, Newey (1997). The first dictionary can approximate the function g_1 sufficiently well so that given $g_2(x)$, it can be estimated in the traditional way and inference on functionals of $g_1(x)$ are reliable. This requires a well-thought-out approximating series provided by the researcher. When the problem of interest is in recovering and performing inference for $g_1(x)$, the second component $g_2(x)$ may be considered a nuisance parameter. In this case, because $g_2(x)$ is not central to inference, added flexibility in choosing the second dictionary by allowing $p_2(x)$ to be high dimensional is permitted. In particular $K_2 >> n$, is allowed. This increased flexibility can potentially increase the robustness of subsequent inference for $g_1(x)$. However, the increased flexibility requires additional structure of $p_2(x)$; the key conditions are sparse approximation. The first sparsity requirement is that there is a small number of components of $p_2(x)$ that adequately approximate the function $g_2(x)$. The second sparsity requirement is that information about functions $h \in \mathcal{G}_1$ conditional on $\mathcal{G}_2$ can be suitably approximated using a small number of terms in $p_2(x)$. The identities of the contributing terms, however, can be unkown to the researcher a priori.

Aside from estimating an entire component of conditional expectation function $g_1(x)$ itself, the structure outlined above will allow estimating certain functionals of $g_1(x)$. Let a be a functional $a(g)$ and suppose that g has a decomposition so that $g(x) = g_1(x) + g_2(x)$ with $a(g) = a(g_1)$. Such functionals include integrals of g_1, weighted average derivatives of $g_1(x)$, evaluation of $g_1(x)$ at a point x^0, and the $\arg\max g_1(x)$. For illustration, suppose that $E[y|x]$ is additively seperable so that $E[y|x] = g_1(x_1) + ... + g_d(x_d)$ for d-dimensional covariate x. Then any functional a of the form $a(g) = E[\partial g/\partial x_1(x)|x]$ satisfies the condition outlined above. Further specialization of the example to a partially linear model $E[y|x] = \alpha x_1 + g(x_2, ..., x_d)$ in which the desired derivative is given by α was explored in Belloni, Chernozhukov and Hansen (2013).

3. Estimation

When the effective number of free parameters is larger than the sample size, model selection or regularization is unavoidable. There are a variety of different model selection techniques available to researchers. A popular approach is via the Lasso estimator given by Tibshirani (1996) which in the context of regression, simultaneously performs regularization and model selection. The Lasso is used in many areas of science and image processesing and has demonstrated good predictive performance. Lasso allows the estimation of regression coefficients even when the sample size is smaller than the number of parameters by

adding to the quadratic objective function a penalty term which mechanically favors regression coefficients that contain zero elements. By taking advantage of ideas in regularized regression, this paper demonstrates that quality estimation of $g_1(x)$ can be attained even when $K_1 + K_2$, the effective number of parameters, exceeds the sample size n. Estimating proceeds by a model selection step that effectively reduces the number of parameters to be estimated.

Estimation of the function $g(x)$ will be based on a reduced dictionary $\tilde{p}(x)$ comprised of a subset of the series terms in $p_1(x)$ and $p_2(x)$. Because the primary object of interest is $g_1(x)$, it is natural to include all terms belonging to $p_1(x)$ in $\tilde{p}(x)$. As mentioned above, this is inclusion is actually unavoidable; the asymptotic normality results require that there is no selection of the terms belonging to $p_1(x)$. [1] The main selection step involves choosing a subset of terms from $p_2(x)$. Suppose that a model selection procedure provides a new dictionary $\tilde{p}_2(x)$, which contains $\tilde{K}_2$ series terms. Each term in $\tilde{p}_2(x)$ is also a term from $p_2(x)$. Then estimation of the function $E[y|x]$ is based on the dictionary $\tilde{p}(x) = (p_1(x), \tilde{p}_2(x)$.

$$\widehat{g}_1(x) = p_1(x)'\widehat{\beta}_1$$

where $(\widehat{\beta}_1', \widehat{\beta}_2')' = (\tilde{P}'\tilde{P})^{-1}\tilde{P}'Y$. Since estimation of $g_2(x)$ is of secondary concern, only the components of $g_2(x)$ that are informative for predicting $g_1(x)$ and y need to be estimated.

There are many candidates for model selection devices in the statistics and econometrics literature. The appropriate choice of model selection methodology can be taylored to the application. In addition to the Lasso, the Scad (Fan (2001)), the BIC, the AIC all feasible. In the exposition of the results, the model selection procedure used will be specifically the Lasso, though results are provided for generic model selection that attains certain performance bounds. Therefore, the next section provides a brief review of issues related to Lasso, especially those that arise in econometric applications.

3.1. **Lasso methods in econometrics.** The following description of the lasso estimator is a review of the particular implementation given in Belloni, Chen, Chernozhukov and Hansen (2013). Consider the conditional expectation $E[y|x] = f(x)$ and assume that $p(x)$ is an approximating dictionary for the function $f(x)$ so that $f(x) \approx p(x)'\beta$. The lasso

[1] An interesting related question, though, is in justifying inference for data dependent K_1 for example, with cross-validation.

estimate for β is defined by

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - p(x_i)'\beta)^2 + \lambda \sum_{j=1}^{K} |\widehat{\Psi}_j \beta_j|$$

where λ and $\widehat{\Psi}_j$ are tuning parameters named the penalty level and the penalty loadings. Belloni, Chen, Chernozhukov and Hansen (2013) provided estimation methodology as well as results guaranteeing performance for the Lasso estimator under conditions which are common in econometrics including heteroskedastic and non-Gaussian disturbances. Tuning parameters are chosen by considerations that balance regularization with bias. For the simple heteroskedastic Lasso above, Belloni, Chen, Chernozhukov and Hansen (2013) recommend setting

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2p), \quad \Psi_j = \sqrt{\sum_{i=1}^{n} p_j(x_i)^2 \epsilon_i^2 / n}$$

with $\gamma \to 0$ sufficiently slowly, and $c > 1$. The choices $\gamma = \log^{-1} n$ and $c = 1.1$ are acceptable. The exact values ϵ_i are unobserved, and so a crude preliminary estimate $\widehat{\epsilon}_i = y_i - \bar{y}$ is used to give $\widehat{\Psi}_j = \sqrt{\sum_{i=1}^{n} p_j(x_i)^2 \widehat{\epsilon}_i^2 / n}$. [2] The preliminary estimates are sufficient for the results below to hold, but the process can be iterated by estimating new residuals using the preliminary estimate. [3]

Lasso performs particularly well relative to some more traditional regularization schemes (eg. ridge regression) under sparsity: the parameter β satisfies $|\{j : \beta_j \neq 0\}| \leqslant s$ for some sequence $s << n$. A feature that has granted Lasso success is that it sets some components of $\widehat{\beta}$ to exactly zero in many cases. and thus serves as a model selection device. The Post-Lasso estimator is defined as the least squares series estimator that considers only terms selected by Lasso (ie terms with nonzero coefficients) in a first stage estimate. Post-Lasso estimation as described above is used as a model selection tool in the subsequent analysis.

3.2. **Post-Double Selection.** The main obstacle in statistical inference after model selection is in attaining robustness to model selection errors. When coefficients are small

[2]This can be iterated as suggested by Belloni, Chen, Chernozhukov and Hansen (2013). The validity of the of the crude preliminary estimate as well as iterative estimates are detailed in the appendix.

[3]A fully data-driven procedure for choosing the penalty level is still unavailable. Cross validation procedures are known to provide relatively low penalty levels so that the regularization event cannot be ensured with high probability

relative to the sample size (ie statistically indistinguishable from zero), model selection mistakes are unavoidable. [4] When such errors are not accounted for, subsequent inference has been shown to be potentially severely misldeading. Difficulties arising from model selection errors under suitable regularity conditions can be overcome through post-double selection, a methodology first proposed by Belloni, Chernozhukov and Hansen (2013). Post-double selection provides an extra measure of robustness by performing two model selection steps before estimating the final model. The basic underlying principal is that regressors misclassified in both model selection steps, and thus wrongly excluded from the model, are those whose omission has negligible effect on inference asymptotically.

To be concrete, in the linear model $E[y_i|d_i, x_i] = \alpha d_i + x_i'\beta$, post double selection considers model selection on two regression equations: (1) the first stage $E[d_i|x_i] = x_i'\beta_{FS}$ and (2) the reduced form $E[y_i|x_i] = x_i'\beta_{RF}$. Estimation of α proceeds by linear regression using those components of x which were selected in one of the above stages. Under appropriate regularity conditions, Belloni, Chernozhukov and Hansen (2013) show that the corresponding $\widehat{\alpha}$ is is consistent and asymptotically Gaussian.

3.3. **Additively Separable Models and Dictionary Selection.** In the additively separable model, the two selection steps are summarized as follows:

(1) *First Stage Model Selection Step* - Select those terms in p_2 which are relevant for predicting terms in p_1.

(2) *Reduced Form Model Selection Step* - Select those terms in p_2 which are relevant for predicting y.

To further describe the first stage selection, consider an operator T on functions that belong to $\mathcal{G}_1$:

$$Th(x) = E[h(x)|\mathcal{G}_2(x)]$$

The operator T measures dependence between functions in the ambient spaces $\mathcal{G}_1, \mathcal{G}_2$ which house the functions g_1, g_2 and the conditioning is understood to be on all function $f \in \mathcal{G}_2$. If the operator T can be suitably well approximated, then the post double selection methodology generalizes to the nonparametric additively separable case. Though it is convenient to consider the general operator T, it is sufficient for estimation purposes to approximate the restriction of T to the subspace spanned by p_1.

[4] Under some conditions, perfect model selection can be attained. For example, beta-min conditions which require that coefficients be either exactly zero or well separate from zero can give perfect model selection

This approximation problem is approached with the Lasso regression. Each component of p_1 is regressed onto the dictionary p_2 giving an approximation for $Tp_{1j}(x)$ as a linear combination of elements $p_2(x)$ for $1 \leqslant k \leqslant K_1$. If this can be done with all p_{1k}, for each $1 \leqslant k \leqslant K_1$, then a linear combination $Tp_1(x)'\beta$ can also be approximated by a linear combination of elements of p_2. The estimation can be summarized with one optimzation problem which is equivalent to K_1 separate Lasso problems. All nonzero components of the solution to the optimization are collected and included as elements of the refined dictionary $\tilde{p}$.

$$\widehat{\Gamma} = \arg\min_{\Gamma} \sum_{j=1}^{K_1} \sum_{i=1}^{n} (p_{1;k}(x_i) - p(x_i)'\Gamma_k)^2 + \lambda^{FS} \sum_{k=1}^{K_1} \sum_{j=1}^{K_2} |\Psi_{jk}^{FS} \Gamma_{kj}|.$$

Note that the estimate $\widehat{\Gamma}$ approximates T in the sense that $(p_2(x)'\widehat{\Gamma})'\beta$ approximates $Tp_1(x)'\beta$. The first stage tuning parameters $\lambda^{FS}, \widehat{\Psi}_{jk}^{FS}$ are chosen similarly to the method outlined above but account for the need to estimate effectively K_1 different regressions. Set

$$\lambda^{FS} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2K_1K_2),$$

$$\Psi_{jk}^{FS} = \sqrt{\sum_{i=1}^{n} p_j(x_i)^2 (p_j(x_i) - Tp_j(x_i))^2 / n}.$$

As before, the Ψ_{jk}^{FS} are not directly observable and so estimates $\widehat{\Psi}_{jk}^{FS}$ are used in their place following the exact method described above .

Running the regression above will yield coefficient estimates of exactly zero for many of the Γ_{kj}. For each $1 \leqslant j \leqslant K_1$ let $\widehat{I}_k = \{j : \Gamma_{kj} \neq 0\}$. Then the first stage model selection produces the terms $\widehat{I}^{FS} = \widehat{I}_1 \cup ... \cup \widehat{I}_{K_1}$.

The reduced form selection step proceeds after the first stage model selection step. For this step, let

$$\widehat{\pi} = \arg\min_{\pi} \sum_{i=1}^{n} (y_i - p_2(x_i)'\pi)^2 + \lambda^{RF} \sum_{j=1}^{K_2} |\widehat{\Psi}_j^{RF} \pi_j|$$

Where the reduced form tuning parameters $\lambda^{RF}, \widehat{\Psi}_{jk}^{RF}$ are chosen according to the method outlined above with

$$\lambda^{RF} = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2K_2),$$

$$\Psi_j^{RF} = \sqrt{\sum_{i=1}^{n} p_j(x_i)^2 (y_i - E[y_i|\mathcal{G}_2(x_i)])^2 / n}.$$

Let $\widehat{I}^{RF} = \{j : \pi_j \neq 0\}$ be the outcome of the reduced form step of model selection.

Considering the set of dictionary terms selected in the first stage and reduced form model selection steps. Let $\widehat{I}$ be the union of all dictionary terms. Then define the refined dictionary by $\tilde{p}(x) = (p_1(x), \{p_{2j}(x)\}_{j \in \widehat{I}})$. Let $\tilde{P}$ be the $n \times (K_1 + |\widehat{I}|)$ matrix with the observations of the refined dictionary stacked. Then $\widehat{\beta} = (\tilde{P}'\tilde{P})^{-1}\tilde{P}'Y$ and $\widehat{g} = \tilde{p}(x)'\widehat{\beta}$. Partitioning $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2)$ leads to the post-double selection estimate of g_1 defined by:

$$\widehat{g}_1 = p_1(x)'\widehat{\beta}_1.$$

There are many alternative model selection devices that can be used in place of lasso. Alternatively, the entire first stage model selection procedure can be done in one step using a group-lasso type penalty which favors a common support in p_2 for approximating all components of p_1. This type of first stage model selection was considered, for example, by Farrell (2013) for a finite number of first stage equations and logistic loss. Further alternatives include square root lasso, SCAD type estimators etc. These examples have all been shown to have good perfomance properties for a single outcome and are expected to exhibit the similar performance as Lasso when the number of regression equations is permitted to grow at a suitalby controlled rate. Therefore, the analysis in this paper focuses on an inference procedure which uses a lasso model selection for each first stage equation.

4. Regularity and Approximation Conditions

In this section, the model described above is written formally and conditions guaranteeing convergence and asymptotic normality of the Post-Double Selection Series Estimator are given.

Assumption 1. *(i) (y_i, x_i) are i.i.d. random variables and satisfy $E[y_i|x_i] = g(x_i) = g_1(x_i) + g_2(x_i)$ with $g_1 \in \mathcal{G}_1$ and $g_2 \in \mathcal{G}_2$ for prespecified classes of functions $\mathcal{G}_1, \mathcal{G}_2$.*

The first assumption specifies the model. The observations are required to be identically distributed, which is stronger than the treatment of i.n.i.d variables given in Belloni, Chernozhukov and Hansen (2013). This can be weakened at the cost of more stringent conditions on the size of the first dictionary using for example the ideas in Andrews 1991.

4.1. Regularity and approximation conditions concerning the first dictionary.

To state the regularity conditions, a few definitions that help characterize the smoothness of target function g_1 and approximating functions p_1. Let$f$ be a function defined on the support $\mathcal{X}$ of x. Define $|f|_d = \sup_{x \in X} \max_{|a| \leq d} \partial^{|a|} f / \partial x^a$. This defines the standard Sobolev

norm. In addition, let $\zeta_d(K_1) = \max_{|a|\leqslant d} \sup_{x\in X} \|\partial^{|a|} p_1(x)/\partial x^a\|$ where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 2. *There is an integer $d \geqslant 0$, a real number $\alpha > 0$, and vectors $\beta_1 = \beta_{1,K_1}$ such that $\|\beta_1\| = O(1)$ and $|g_1 - p_1'\beta_1|_d = O(K_1^{-\alpha})$ as $K_1 \to \infty$.*

Assumption 2 is standard in nonparametric estmiation. It requires that the dictionary p_1 can approximate g_1 at a prespecified rate. Values of d and α can be derived for particular classes of functions. Newey (1997) gives approximation rates for several leading examples, for instance orthogonal polynomials, regression splines, etc.

Assumption 3. *For each K_1, the smallest eigenvalue of the matrix*

$$E\left[(p_1(x) - Tp_1(x))(p_1(x) - Tp_1(x))'\right]$$

is bounded uniformly away from zero in K_1. In addition, there is a sequence of constants $\zeta_0(K)$ satisfying $\sup_{x\in\mathcal{X}} \|p_1(x)\| \leqslant \zeta_0(K_1)$ and $\zeta_0(K_1)^2 K_1/n \to 0$ as $n \to \infty$.

The next condition is a direct analogue of a combination of Assumption 2 from Newey (1997) and the necessary and sufficient conditions for estimation of partially linear models from Robinson (1988). Requiring $E\left[(p_1(x) - Tp_1(x))(p_1(x) - Tp_1(x))'\right]$ to have uniformly bounded away from zero eigenvalues is an identifiability condition. It is an analogue of the standard condition that $E[p(x)p(x)']$ have eigenvalues bounded away from zero specialized to the residuals of $p_1(x)$ after conditioning on $\mathcal{G}_2(x)$. The second statement of Assumption 2 is a standard regularity condition on the first dictionary.

4.2. **Sparsity Conditions.** The next assumptions concern sparsity properties surrounding the second dictionary. As outlined above, sparsity will be required along two dimensions in the second dictionary: both with respect to the outcome equation (1) and with respect to the functional T. Consider a sequence $s = s_n$ that controls the number of nonzero coefficients in a vector. A vector X is $s-$sparse if $|\{j : X_j \neq 0\}| \leqslant s$. The following give formal restrictions regarding the sparsity of the outcome equation relative to the second approximating dictionary as well as a sparse approximation of the operator T described above.

Assumption 4. *Sparsity Conditions. There is a sequence $s = s_n$ and $\phi = s\log(\max\{K_1K_2, n\})$ such that*

(i) Approximate sparsity in the outcome equation. There is a sequence of vectors $\beta_2 = \beta_{2,K_2}$ that are s-sparse and the approximation $\sqrt{\sum_{i=1}^n (g_2(x_i) - p_2(x_i)'\beta_2)^2/n} := \xi_0 = O_P(\sqrt{\phi}/n)$ holds.

(ii) Approximate sparsity in the first stage. There are $s-$sparse $\Gamma_k = \Gamma_{k,K_2}$ such that $\max_{k \leqslant K_1} \sqrt{\sum_{i=1}^n \left(E[p_{1k}(x)|\mathcal{G}_2(x)] - p_2(x)'\Gamma_k\right)^2 / n} := \xi_{FS} = O_P(\sqrt{\phi/n})$.

Note: The assumption above imposes no conditions on the sparsity s. This is postponed until Assumption 6. The conditions listed in Assumption 6 will require that $K_1^{3/2}\phi n^{-1/2} \to 0$

The first statement requires that the second dictionary can approximate g_2 is a small number of terms. The restriction on the approximation error follows the convention used by Belloni, Chen, Chernozhukov and Hansen (2013). The average squared approximation error from using a sparse β_2 must be smaller than the conjectured estimation error when the support of the correct small number of terms is known. The second statement generalizes the the first approximate sparsity requirement. It requires that each component of the dictionary p_1 can be approximated by a linear combination of a small set of terms in p_2. The second statement is substantive because it requires for each k that there be a relatively small number of elements of $p_2(x)$ which can adequately the conditional expectation of each term in $p_1(x)$. Finally, the third condition formalizes the rate at which the sparsity index can increase relative to the sample size. The assumption is substantive and implies that favorable estimation results are only guaranteed if the number of relevant series terms in $p_2(x)$ is small in comparison to the sample size. In addition, if more terms are required in $p_1(x)$ to approximate $g_1(x)$, so that K_1 is inflated, then the restrictions on the sparsity are even more stringent. This implies that not only do y and p_1 require a sparse approximation with $p_2(x)$, but also that $g_1(x)$ is particularly smooth or well behaved.

4.3. **Regularity conditions concerning the second dictionary.** The following conditions restricts the sample Gram matrix of the second dictionary. A standard condition for nonparametric estimation is that for a dictionary P, the Gram matrix $P'P/n$ eventually has eigenvalues bounded away from zero uniformly in n with high probability. If $K_2 > n$, then the matrix $P_2'P_2/n$ will be rank deficient. However, it is sufficient that only small submatrices of $P_2'P_2/n$ have the desired property. In the sparse setting, it is convenient to define the following sparse eigenvalues of a positive semi-definite matrix M:

$$\varphi_{\min}(m)(M) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}, \quad \varphi_{\max}(m)(M) := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}$$

In principal, the sparse eigenvalues of the sample Gram matrix are observed, however, explicitly calculating them for each $m < n$ is computationally prohibitive. However, under many classes of data generating processess, the sparse eigenvalues are known to be bounded away from zero with high probability. See Bickel, Ritov and Tsybakov (2009), Zhou (2009)

for explicit examples of dgps for which sparse eigenvalue conditions hold. In this paper, favorable behavior of restricted eigenvalues is taken as a high level condition. Impose the following simple assumption.

Assumption 5. *There is a sequence $\ell_n \to \infty$ and constants $\kappa'' > \kappa' > 0$ such that the sparse eigenvalues obey*

$$\kappa' \leqslant \varphi_{\min}(\ell_n s K_1)(P_2'P_2/n) \leqslant \varphi_{\max}(\ell_n s K_1)(P_2'P_2/n) \leqslant \kappa''.$$

4.4. **Moment Conditions.** The final conditions are moment conditions which ensure good performance of the Lasso as a model selection device. They allow the use of moderate deviation results given in Jing Shao and Wang (2003) which ensures good performance of Lasso under non-Gaussian and heteroskedastic errors. In addition, the moment conditions are needed in order to guarantee that the the validity of the plug in variance estimator. Belloni, Chernozhukov and Hansen (2013) discuss plausibility of these types of moment conditions for various models for the case $K_1 = 1$. For common approximating dictionaries for a single variable, the condition can be readily checked in a similar manner. This is only one possible set of moment conditions.

Assumption 6. *Let $\epsilon = y - g(x)$. For each $k \leqslant K_1$ let $W_k = p_{1k}(x) - Tp_{1k}(x)$. Let $q > 4$. Let c, C be constants that do not depend on n and can take different values in each occurance. The following moment conditions are satisfied:*

(i) For each j, $E[|p_{1j}(x_i)|^q]$. $c \leqslant E[\epsilon_i^2|x_i] \leqslant C$. $E[\epsilon_i^4|x_i] \leqslant C$.

(iii) $E[|\epsilon_i|^q] + E[y_i^2] + \max_{j \leqslant K_2}\{E[p_{2j}(x_i)^2 y_i^2] + E[|p_{2j}(x_i)|^3|\epsilon|^3] + E[p_{2j}(x_i)]^{-1}\} \leqslant C$

(iv) For each $k \leqslant K_1$, $E[|W_k|^q] + E[p_{1k}(x)^2]$
$+ \max_{j \leqslant K_2}\{E[p_{2j}(x_i)^2 p_{1k}(x_i)^2] + E[|p_{2j}(x_i)|^3|W_{k_i}|^3] + E[p_{2j}(x_i)]^{-1}\} \leqslant C.$

(v) $\log^3 K_2/n = o(1)$.

(vi) $\max_{i,j} p_{2j}(x_i)^2\phi/n = o_P(1)$

(vii) $\max_{k \leqslant K_1, j \leqslant K_2} |\sum_i p_{2j}(x_i)^2(W_{ki}^2 + \epsilon_i^2)/n - E[p_{2j}(x)^2(W_{ki}^2 + \epsilon_i^2)]| = o_P(1)$

(viii) $K_1^{3/2}\phi n^{-1/2+2/q} \to 0$.

4.5. **Global Convergence.** The first result is a preliminary result which gives bounds on convergence rates for the estimator $\widehat{g}_1$. Though they are of interest in their own right in that they develope a direct comparison of pervious methods with high dimensional methods, they are used in the course of the proof of Theorem 2 which is the main inferential result.

The proposition is a direct analogue of the rates given in Theorem 1 of Newey (1997). This is a demonstration that identical to those encountered in classical nonparametric estimation can often be recovered when considering an isolated component of a very high dimensional problem.

Proposition 1. *Under assumptions listed above, the post double selection estimates for the function g satisfy*

$$\int (g_1(x) - \widehat{g}_1(x))^2 dF(x) = O_p(K_1/n + K_1^{-2\alpha})$$
$$|\widehat{g}_1 - g_1|_d = O_P(\zeta_d(n)\sqrt{K_1}/\sqrt{n} + K_1^{-\alpha}).$$

The result is stated for congergence of $\widehat{g}_1$. A similar result is expected to hold for the second component, $\widehat{g}_2$ and therefore also $\widehat{g}$. However, the asymptotic normality results will not hold for g_2 and therefore, the Theorem is stated concerning g_1 only.

5. Inference and asymptotic normality

In this section, formal results concerning inference are stated. Proofs of the theorems are provided in the appendix. The theorem concerns estimators which are asymptotically normal and is the main consideration of the paper. In particular, consider estimation of a functional a on the class of functions $\mathcal{G}_1$. The quantity of interest, $\theta = a(g_1)$, is estimated by

$$\widehat{\theta} = a(\widehat{g}_1).$$

The following assumptions on the functional a are imposed. They are regularity assumptions that imply that a attains a certain degree of smoothness. For example, they imply that a is Fréchet differentiable.

Assumption 7. *Either (i) a is linear over $\mathcal{G}_1$; or (ii) for d as in the previous assumption, $\zeta_d(K_1)^4 K_1^2/n \to 0$. In addition, there is a linear function $D(f, \tilde{f})$ that is linear in f and such that for some constants $C, \nu > 0$ and all $\bar{f}, \tilde{f}$ with $|\tilde{f} - g_1|_d < \nu$, $|\bar{f} - g_1|_d < \nu$, it holds that $\|a(f) - a(\tilde{f}) - D(f - \tilde{f}; \tilde{f})\| \leqslant C(|f - \tilde{f}|_d)^2$ and $\|D(f; \tilde{f}) - D(f; \bar{f})\| \leqslant L|f|_d|\tilde{f} - \bar{f}|_d$.*

The function D is related to the functional derivative of a. The following assumption imposes further regularity on the continuity of the derivative. For shorthand, let $D(g) = D(g; g_0)$.

Assumption 8. *Either (i) a is scalar, $|D(g_1)| \leqslant C|g|_d$ for d described above. There is $\bar{\beta}_1$ dependent on K_1 such that for $\bar{g}_1(x) = p_1(x)'\bar{\beta}_1$, it holds that $E[\bar{g}_1(x)^2] \to 0$ and*

$D(\bar{g}_1) \geqslant C > 0$; *or (ii) There is* $v(x)$ *with* $E[v(x)v(x)']$ *finite and nonsingular with* $D(g_1) = E[v(x)g_1(x)]$ *and* $D(p_{1k}) = E[v(x)p_{1k}(x)]$ *for every* k. *There is* $\tilde{\beta}_1$ *so that* $E[\|v(x) - p_1(x)'\tilde{\beta}\|^2] \to 0$.

In order to use $\widehat{\theta}$ for inference on θ, an approximate expression for the variance $var(\widehat{\theta})$ is necessary. As is standard, the expression for the variance will be approximated using the delta method. An approximate expression for the variance of the estimator $\widehat{\theta}$ therefore requires an appropriate derrivative of the function a, (rather, an estimate). Let A denote the derivatives of the functions belonging to the approximating dictionary, $A = (D(p_{11}, ..., D(p_{1K_1})'$. Let $\widehat{A} = \frac{\partial a(p_1(x)'b}{\partial b}(\widehat{\beta}_1)$ provided that $\widehat{A}$ exists. Let

The approximate variance, given by the delta method is given by $V = V_{K_1}$:

$$
\begin{aligned}
V &= AQ^{-1}\Sigma Q^{-1}A \\
Q &= E[(p_1(x) - Tp_1(x))(p_1(x) - Tp_1(x))'] \\
\Sigma &= E[(p_1(x) - Tp_1(x))(p_1(x) - Tp_1(x))'(y - g(x))^2]
\end{aligned}
$$

These quantities are unobserved but can be estimated:

$$
\begin{aligned}
\widehat{V} &= \widehat{A}\widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}\widehat{A} \\
\widehat{Q} &= \sum_{i=1}^{n}(p_1(x_i) - \widehat{p}_1(x_i))(p_1(x_i) - \widehat{p}_1(x_i))'/n \\
\widehat{\Sigma} &= \sum_{i=1}^{n}(p_1(x_i) - \widehat{p}_1(x_i))(p_1(x_i) - \widehat{p}_1(x_i))'(y - \widehat{g}(x_i))^2/n
\end{aligned}
$$

The estimates $\widehat{p}_1$ are defined as componentwise least squares projections. For each $k \leqslant K_1$, $\widehat{p}_{1k}$ is the orthogonal projection of p_{1k} onto the components of p_2 that belong to $\widehat{I}$. Them $\widehat{V}$ is used as an estimator of the asymptotic variance of $\widehat{\theta}$ and assumes a sandwhich form.

The next result is the main result of the paper. It establishes the validity of standard inference procedure after model selection as well as validity of the plug in variance estimator.

Theorem 1. *Under the Assumptions 1-7 and Assumption 8(i), and if in addition* $\sqrt{n}K^{-\alpha} \to 0$ *then* $\widehat{\theta} = \theta + O_P(\zeta_d(K_1)\sqrt{n})$ *and*

$$
\sqrt{n}V^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0,1), \qquad \sqrt{n}\widehat{V}^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0,1)
$$

If Assumptions 1-7 and Assumption 8(ii) hold and $d = 0$ *and* $\sqrt{n}K^{-\alpha} \to 0$ *then for* $\bar{V} = E[v(x)v(x)'var(y|x)]$*, the following convergences hold.*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \bar{V}), \qquad \|\widehat{V} - \bar{V}\| \xrightarrow{p} 0$$

This establishes the validity of standard inference for functionals after selection of series terms. Note that under assumption 8(i) the $\sqrt{n}$ rate is not achieved because the the functional a does not have a mean square continuous derivative. By contrast, Assumption 8(ii) is sufficient for $\sqrt{n}$-consistency. Conditions under which the particular assumptions regarding the approximation of g_1 hold are well known. For example, conditions on K_1 for various common approximating dictionaries including power series or regression splines etc follow those directly derived in Newey (1997). Asymptotic normality of these types of estimates under the high dimensional additively separable setting should therefore be viewed as a corollary to the above result.

Consider one example with the functional of interest being evaluation of g_1 at a point x^0: $a(g_1) = g_1(x_0)$. In this case, a is linear and $D(\bar{g}) = \bar{g}(x_0)$ for all functions $\bar{g}$. This particular example does not attain a $\sqrt{n}$ convergence rate provided there is a sequence of functions g_{1K} in the linear span of $p_1 = p_1^K$ such that $E[g_{1K}(x)^2]$ converges to zero but $g_{1K}(x_0)$ is positive for each K. Another example is the weighted average derivative $a(g_1) = \int w(x)\partial g_1(x)/\partial x$ for a weight function w which satisfies regularity conditions. For example, the theorem holds if w is differentiable , vanishes outside a compact set, and the density of x is bounded away from zero wherever w is positive. In this case, $a(g_1) = E[v(x)g_1(x)]$ for $v(x) = -f(x)^{-1}\partial w(x)/\partial x$ by a change of variables provided that x is continuously distributed with non vanishing density f. These are one possible set of sufficient conditions under which the weighted average derivative does achieve $\sqrt{n}$-consistency.

6. Simulation study

The results stated in the previous section suggest that post double selection type series estimation should exhibit good inference properties for additively separable conditional expectation models when the sample size n is large. The following simulation study is conducted in order to illustrate the implementation and performance of the outlined procdure. Results from several other candidate estimators are also calculated to provide a comparison between the post-double method and other methods. Estimation and inference for two functionals of the conditional expectation function are considered. Two simulation designs are considered. In one design, the high dimensional component over which model selection

is performed is a large series expansion in four variables. In the other design, the high dimensional component is a linear function of a large number of different covariates.

6.1. **Low Dimensional Additively Separable Design.** Consider the following model of six continuous variables $x_1, ..., x_5$ of form:

$$E[y|x] = E[y|x_1, ..., x_5] = g_1(x_1) + g_2(x_2, ..., x_5)$$

The appearence of the term $g_2(x_2, ..., x_5)$ can be problematic for standard nonparametric regression because g_2 is an unspecified function of 5 variables and so the dimensionality of the problem becomes a burdon. The objective is to estimate a population averge derivative, $\theta^{(1)}$, and a function evaluation, $\theta^{(2)}$ given by

$$\theta^{(1)} = a^{(1)}(g_1) = \int_{\mathcal{X}} \partial g/\partial x_1 dF(x_1), \quad \theta^{(2)} = a^{(2)}(g_1) = g_1(E[x_1]).$$

The average derivative is integrated over the middle 50-percent of the values assumed by x_1. The integration is performed over the central two quartiles in order to avoid edge effects. According to the theorem, both parameters can be estimated compared to a Gaussian reference distribution for testing. The true function of interest, g_1, used in the simulation is given by the functional form:

$$g_1(x_1) = (x_1 - 3) - (x_1 - 3)^2/12 + 2\frac{\exp(x_1 - 3)}{1 + \exp(x_1 - 3)} + C$$

with the constant C in the expression for g_1 is defined to ensure that $g_1(0) = 0$. g_1 is comprised of a quadratic function and a logistic function of x_1. Ex post, the function is simple, however, for the sake of the simulation, knowledge of the logistic form is assumed unkown. The second component is given by

$$g_2(x_2, ..., x_6) = -6(x_2 - x_2x_3 + x_3) + 6\frac{\exp(x_4 + x_5)}{1 + \exp(x_4 + x_5)}$$

The second function g_2 is similar, being defined by a combination of a logistic function and a quadratic function. The logistic part can potentially require many interaction terms unkown in advance to produce an accurate model. The component functions g_1 and g_2 will be used throughout the simulation. The remaining parameters, eg. dictating the data generating processes for $x_1, ..., x_5$ will be changed accross simulation to give an illustration of performance accross different settings.

The covariates and outcome are drawn as follows. The marginal distribution of the regressors $x_2, ..., x_5$ iis set at $N(0, 1)$; their correlations are set to $\text{corr}(x_j, x_k) = \left(\frac{1}{2}\right)^{-|j-k|}$. The variable of interest, x_1 is determined by $x_1 = 3 + (x_2 + ... + x_5) \cdot + v \cdot \sigma_v$ with $v \sim N(0, 1)$ independent of $x_2, ...x_5$. The structural errors $\epsilon = y - E[y|x]$ are drawn $N(0, 1) \cdot \sigma_\epsilon$ independent of x. Several simulations of this model are conducted by varying the sample

size n, the dependence between x_1 and the remaining regressors $(x_2, ..., x_5)$, as well as the size of the residual errors ϵ. The sample size is set to either $n = 500$ or $n = 800$. The dependence between x_1 and the remaining covariates is dictated by σ_v. To capture high dependence between the covariates, the value $\sigma_v = 3$ is used, and to capture low dependence, $\sigma_v = 5$ is used. Finally, the variability of the structural shocks are set to $\sigma_\epsilon = 3$ and $\sigma_\epsilon = 5$.

Estimation is based on a cubic spline interpolation for the first dictionary. When the sample size is $n = 500$, knot points are given at $x_1 = -2, 2, 6, 10$. When the sample size is $n = 800$, additional knot points are used at $x_1 = 0, 4, 8$. The second dictionary is comprised of cubic splines in each variable with knots at $x_j = -1, 0, 1$. When $n = 500$, interactions of the variables are allowed for only the linear and quadratic terms. When $n = 800$, the spline terms are allowed to have interactions. At most 3 terms from each marginal dictionary for the x_j are allowed to be interacted. This gives $K_2 = 170$, $K_2 = 640$ for the two sample sizes.

In addition to the post-double lasso based model, several alternative estimates are calculated for comparison. Two standard series estimator which are designed to approximate the function $g = g_1 + g_2$ is given. The first is based on a series approximation which uses polynomials up to order two. The second series estimator is allows general polynomials up to order two but also allows powers of each individual variable up to order 4 (ie no higher order interaction terms). Second, a single step selection estimator is provided. The single step estimator is done by performing a first stage lasso on the union of the two dictionaries, then reestimating coefficients of the remaining dictionary terms in the second stage. Finally, an infeasible estimator is provided, where estimation proceeds as standard series estimation given the dictionary p_1 and as if $g_2(x_i)$ where known for each i.

Because the true population distribution of $dF(x_1)$ is not known exactly, an estimate is used for calculating $\widehat{a}^{(1)}$ and $\widehat{a}^{(2)}$. In particular, $\widehat{a}^{(1)}$ is given by the derivative of $\widehat{g}_1$ integrated against the empirical distribution $d\widehat{F}(x_1)$. $\widehat{a}^{(1)} = \int \partial \widehat{g}_1/\partial x_1(t) d\widehat{F}(t) = \sum_{i=1}^n \partial \widehat{g}_1/\partial x_1(x_{1i})$. Similarly, $\widehat{a}^{(2)} = \widehat{g}_1(\int x_1 d\widehat{F}(x_1))$. Given the true function g_1, and the true distribution $dF(x_1)$, the true value for $\theta^{(1)}, \theta^{(2)}$ is apprixmately $\theta_0^{(1)} = 4.20$ and $\theta^{(2)} = 10.32$. These value will be used for hypothesis testing in the simulation.

Results for estimating $\theta^{(1)}$ and $\theta^{(2)}$ are based on 500 simulations for each setting described earlier. For each estimator, the median bias, median absolute deviation, and rejection probability for a 5-percent level test of $H_0^{(1)} : \theta^{(1)} = \theta_0^{(1)}$ or $H_0^{(2)} : \theta^{(2)} = \theta_0^{(2)}$ are presented. Results for estimating $\theta^{(1)}$ are presented in Table 1. In each of the simulation, estimates of θ based on post double selection exhibit small median absolute deviation relative to

the competing estimates. With the exception of the infeasible estimates the post double selection estimates are also the only estimates which exhibit reasonable rejection frequencies consistently accross all settings. Results for estimation of $\theta^{(2)}$ are reported in Table 2. The results are qualitatively similar to those for $\theta^{(1)}$. The only reasonable rejection frequencies are obtained with the post-double selection. A small amount of size distortion can be seen as rejection frequencies are closer to 10-percent in most simulations. The distortion in the post double estimator matches that in the infeasible estimator suggesting that they are driven by bias in approximating a nonlinear function with a small bais rather than a consequence of model selection.

6.2. **High Dimensional Additively Separable Design.** In this design, a high dimensional setting is considered:

$$E[y|x] = E[y|x_1, ..., x_p] = g_1(x_1) + g_2(x_2, ..., x_p).$$

The model now depends on p parameters which is allowed to change with the sample size. For $n = 500$, $p = 400$ is used, while for $n = 800$, $p = 640$ is used. The variables $x_2, ..., x_p$ are drawn marginally from $N(0, 1)$ with correlation structure $\text{corr}(x_j, x_k) = \left(\frac{1}{2}\right)^{-|j-k|}$. The target function $g_1(x_1)$ remains the same as before. The function $g_2(x_2)$ is given by $g_2(x_2, ..., x_p) = \sum_j .7^{-j} x_j$. The dependence between x_1 and $x_2, ..., x_p$ is defined with $x_1 = 3 + \sum_j .7^{-j} x_j + v \cdot \sigma_v$. The specifications for v, σ_v, ϵ, and σ_ϵ are the same as they were in the low dimensional example. Estimation is again based on a cubic spline interpolation for the first dictionary. When the sample size is $n = 500$, knot points are given at $x_1 = -2, 2, 6, 10$. When the sample size is $n = 800$, additional knot points are used at $x_1 = 0, 4, 8$. The second dictionary comprises of the variables $x_2, ...x_p$.The simulation evaluates the performance of the estimator when the goal is to control for many distinct possible sources of confounding. Results are recorded in Table 3 for sample size $n = 500$ and Table 4 for $n = 800$. In this simulation, the single selection and infeasible estimators are defined as they were in the low dimensional simulation. The series estimator simply uses the entire set of controls in estimation. The results are qualitatively similar to those for the first design. The only reasonable rejection frequencies are obtained with the post-double selection.

7. Conclusion

This paper provides convergence rates and inference results for series estimators with a high dimensional component. In models that admit an additively seperable form, an single

component can be estimated with standard rates customary in nonparametric estimation, even when then number of remaining terms is large relative to the sample size. Restrictions on the first dictionary are exactly like those standard in nonparametric series estimation.

8. References

Andrews, D.W.K., and Y.J. Whang (1995): "Additive interactive regression models: circumvention of the curse of dimensionality." Econometric Theorey 6, 466-479.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2013): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica

Belloni, A., V. Chernozhukov, and C. Hansen (2011): "Inference for High-Dimensional Sparse Econometric Models," Advances in Economics and Econometrics. 10th World Congress of Econometric Society.

Belloni, A., V. Chernozhukov, and C. Hansen (2013): "Inference on treatment effects after selection amongst high-dimensional controls," Advances in Economics and Econometrics. 10th World Congress of Econometric Society.

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

Bhlmann, P. (2013). "Statistical significance in high-dimensional linear models." Bernoulli 19, 1212-1242.

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

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.

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

de la Pena, 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.

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.

Farrell, M.H., (2013): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," Working Paper.

Gautier, E., and A. Tsybakov (2011): "High-dimensional Instrumental Variables Rergession and Confidence Sets," arXiv:1105.2454v2 [math.ST]

Huang, J., J. L. Horowitz, and F. Wei (2010): "Variable selection in nonparametric additive models," Ann. Statist., 38(4), 2282-2313.

Leeb, H., and M. Potshcer (2008): "Can one estimate the unconditional distribution of post-model-selection estimators?" Econometric Theorey 24(2) 388-376.

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

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

Potscher, B. M. (2009): "Confidence sets based on sparse estimators are necessarily large," Sankhya, 71(1, Ser. A), 1-18.

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

Stone, C.J., (1982): "Optimal global rates of convergence for nonparametric regression." Annals of Statistics, 13, 689-705.

Stone C.J., (1985): "Additive regression and other nonparametric models." Annals of Statistics 13, 689-705.

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.

Zhou, S. (2009): "Restricted eigenvalue conditions on subgaussian matrices," ArXiv:0904.4723v2.

9. Appendix: Proofs of the Main Results

9.1. **Additional notation used in proofs.** In the course of the proofs, the following notation will be used. Let $\widehat{I}$ be the full set of series terms chosen in the final estimation coming from the dictionary p_2. $\widehat{I}$ is given by $\widehat{I} = \widehat{I}_0 \cup \widehat{I}_{R.F.} \cup \widehat{I}_1 \cup, ..., \cup I_{K_1}$. Define for any subset $J \subset [p]$, $P_2[J]$ to be the corresponding set of selected dictionary elements. Let b be the least squares coefficient for the regression of any vector U on $P_2[J]$ so that $b = b(U;J) = (P_2[J]'P_2[J])^{-1}P_2[J]'U$. Let $\mathscr{P}_{\widehat{I}} = P_2[\widehat{I}](P_2[\widehat{I}]'P_2[\widehat{I}])^{-}P_2[\widehat{I}]$ be the sample projection onto the space spanned by $P_2[\widehat{I}]$. Let $\mathscr{M}_{\widehat{I}} = I_n - \mathscr{P}_{\widehat{I}}$ be projection onto the corresponding orthogonal space. Let $\widehat{Q} = P_1'\mathscr{M}_{\widehat{I}}P_1/n$. Let $Q = E[(p_1(x) - E[p_1(x)|\mathcal{G}_2])(p_1(x) - E[p_1(x)|\mathcal{G}_2])']$. Similarly, decompose $P_1 = m + W$ where $m = E[P_2|\mathcal{G}_2(X)]$. Let $\bar{Q} = W'W/n$. Let $\|\cdot\|$ denote Euclidean norm when applied to a vector and the matrix norm $\|A\| = \sqrt{\text{tr}A'A}$ when applied to a square matrix. Let $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote L_1 and L_∞ norms. Let $\xi_{FS} = \max_k \sqrt{(m_k - P_2\Gamma_k)'(m_k - P_2\Gamma_k)/n}$ and $\xi_{RF} = \sqrt{(G_1 + G_2 - P_2\pi)'(G_1 + G_2 - P_2\pi)/n}$ be the approximation error in the first stage and reduced form.

9.2. **Proof of Proposition.** Begin by establishing the claim $\|\widehat{Q} - Q\| \xrightarrow{p} 0$ by bounding each of the following terms seperately: $\|\widehat{Q} - Q\| = \|\widehat{Q} - \bar{Q} + \bar{Q} - Q\| \leqslant \|\widehat{Q} - \bar{Q}\| + \|\bar{Q} - Q\|$. The argument in Theorem 1 of Newey (1997), along with the fact that $\sup_{x\in\mathcal{X}} \|p_1(x) - Tp_1(x)\| \leqslant 2\sup_{x\in\mathcal{X}} \|p_1(x)\|$ gives the bound $\|\bar{Q} - Q\| = O_P(\zeta_0(K_1)K_1^{1/2}/\sqrt{n})$. Next bound $\|\widehat{Q} - \bar{Q}\|$. Using the decomposition, $P_1 = m + W$, write $\widehat{Q} = (m+W)'\mathscr{M}_{\widehat{I}}(m+W)/n = W'W/n - W'(I_n - \mathscr{M}_{\widehat{I}})W/n + m'\mathscr{M}_{\widehat{I}}m/n + 2m'\mathscr{M}_{\widehat{I}}W/n$. By triangle inquality, $\|\bar{Q} - \widehat{Q}\| \leqslant \|W'\mathscr{P}_{\widehat{I}}W/n\| + \|m'\mathscr{M}_{\widehat{I}}m/n\| + \|2m'\mathscr{M}_{\widehat{I}}W/n\|$. Bounds for each of the three previous terms are established in Lemma 4 giving $\|\bar{Q} - \widehat{Q}\| = O_P(K_1\phi/n)$.

Since Q has minimal eigenvalues bounded from below by assumption, it follows that $\widehat{Q}$ is inveritble with probability approaching 1. Consider the event $\mathscr{L} = \{\lambda_{\min}(\widehat{Q}) > \lambda_{\min}(Q)/2\}$. By reasoning identical to that given in Newey (1997), it follows that $1_{\mathscr{L}}\|\bar{Q}^{-1}W'\epsilon/n\| = O_P(\sqrt{K_1}/\sqrt{n})$ and $1_{\mathscr{L}}\|\bar{Q}^{-1}W'(G_1 - P_1\beta_1)/n\| = O_P(K_1^{-\alpha})$. To proceed, it is required to obtain analogous bounds for $1_{\mathscr{L}}\|\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n\|$ and $1_{\mathscr{L}}\|\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 - P_1\beta_1)/n\|$. Note that

$$1_{\mathscr{L}}\|\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n - \bar{Q}^{-1}W'\epsilon/n\| \leqslant 1_{\mathscr{L}}\|(\widehat{Q}^{-1} - \bar{Q}^{-1})W'\epsilon/n\| + 1_{\mathscr{L}}\|\bar{Q}^{-1}(W' - P_1'\mathscr{M}_{\widehat{I}})\epsilon/n\|.$$

Consider the first term above. $1_{\mathscr{L}}\|(\widehat{Q}^{-1} - \bar{Q}^{-1})W'\epsilon/n\| \leqslant 1_{\mathscr{L}}\lambda_{\max}(\widehat{Q}^{-1} - \bar{Q}^{-1})\|W'\epsilon/n\| = O_P(\sqrt{K_1/n})O_P(\zeta_0(K_1)\sqrt{K_1}/\sqrt{n}) = O_P(\sqrt{K_1/n})$. Second, $\|\bar{Q}^{-1}(W' - P_1\ \mathscr{M}_{\widehat{I}})\epsilon/n\| \leqslant \lambda_{\max}(\bar{Q}^{-1})\|(W' - P_1'\mathscr{M}_{\widehat{I}})\epsilon/n\| = O_P(1)\|m'\mathscr{M}_{\widehat{I}}\epsilon/n\| = O_P(\sqrt{K}_1\phi/\sqrt{n})$ by Lemma 4. Also,

$1_{\mathscr{L}}\|\bar{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 - P_1\beta_1)/n\| = O_P(K_1^{-\alpha})$ by the same argument as $1_{\mathscr{L}}\|\bar{Q}^{-1}W'(G_1 - P_1\beta_1)/n\| = O_P(K_1^{-\alpha})$.

Also, $1_{\mathscr{L}}\|\widehat{Q}^{-1}P'\mathscr{M}_{\widehat{I}}G_2/n\| = O_P(1)\|P'\mathscr{M}_{\widehat{I}}G_2/n\| = O_P(\zeta_0(K_1)\sqrt{K_1\phi/n} + \sqrt{K_1\phi^2/ns} + \sqrt{K}_1K^{-\alpha}\sqrt{\phi/s})/\sqrt{n} + O_P(\sqrt{K_1\phi}K_1^{-\alpha} + \sqrt{K_1}\sqrt{K_1^2\phi^2/n}\sqrt{\phi/n})/\sqrt{n}$ by triangle inequality and Lemma 4(iv) and 4(v). This reduces to $o_P(\sqrt{K_1/n} + K^{-\alpha})$.

To show the proposition, bound the difference $\widehat{\beta}_1 - \beta_1$. Note that $1_{\mathscr{L}}(\widehat{\beta}_1 - \beta_1) = 1_{\mathscr{L}}\widehat{Q}^{-1}P_1'M_{\widehat{I}}(y - G_1 - G_2)/n + 1_{\mathscr{L}}\widehat{Q}^{-1}P'M_{\widehat{I}}(G_1 - P_1\beta_1)/n - 1_{\mathscr{L}}\widehat{Q}^{-1}P'M_{\widehat{I}}G_2/n$. Triangle inequality and bounds described above give $1_{\mathscr{L}}\|\widehat{\beta}_1 - \beta_1\| \leqslant 1_{\mathscr{L}}\|\widehat{Q}^{-1}P'M_{\widehat{I}}\epsilon/n\| + 1_{\mathscr{L}}\|\widehat{Q}^{-1}P_1'M_{\widehat{I}}(G_1 - P_1\beta_1)/n\| + 1_{\mathscr{L}}\|\widehat{Q}^{-1}P_1'M_{\widehat{I}}G_2/n\| = O_p(K^{1/2}/\sqrt{n} + K^{-\alpha})$. The statement of the proposition follows from the bound on $\widehat{\beta}_1 - \beta_1$.

9.3. **Proof of Theorem.** The proof follows the outline set by Newey (1997) but accounts for model selection considerations. Let $F = V^{-1/2}$ and $\bar{g}_1 = p_1(x)'\beta_2$ and decompose the quantity $1_{\mathscr{L}}\sqrt{n}F[a(\widehat{g}_1) - a(g_1)]$ by

$$1_{\mathscr{L}}\sqrt{n}F[a(\widehat{g}_1)-a(g_1)] = 1_n\sqrt{n}F[a(\widehat{g}_1)-a(g_1)+D(\widehat{g}_1)-D(g_1)+D(\bar{g}_1)-D(g_1)+D(\widehat{g}_1)-D(\bar{g}_1)].$$

By arguments given in the proof of Theorem 2 in Newey (1997), $1_{\mathscr{L}}|\sqrt{n}F[D(\bar{g}_1) - D(g_1)]| \leqslant C\sqrt{n}K_1^{-\alpha}$. In addition, bounds on $|\widehat{g}_1 - g_1|_d$ given by the proposition imply that $|\sqrt{n}F[a(\widehat{g}_1) - a(g_1) - D(\widehat{g}_1) + D(g_1)| \leqslant L\sqrt{n}|\widehat{g}_1 - g_1|_d^2 = O_P(L\sqrt{n}\zeta_d(K_1)(\sqrt{K}_1/\sqrt{n} + K_1^{-\alpha} + \sqrt{s}/\sqrt{n})^2) \to 0$. It remains to be shown that $1_{\mathscr{L}}\sqrt{n}F[D(\widehat{g}_1) - D(\bar{g}_1)]$ satisfies an appropriate central limit theorem. Note that $D(\widehat{g}_1)$ can be expanded

$$\begin{aligned} D(\widehat{g}_1) &= D(p_1(x)'\widehat{\beta}_1) = D(p_1(x)'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}y) \\ &= D(p_1(x)'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 + G_2 + \epsilon)) = D(p_1(x))'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 + G_2 + \epsilon) \\ &= A'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 + G_2 + \epsilon) = A'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_1 + A'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_2 + A'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon \end{aligned}$$

Using the above expansion and $D(\bar{g}_1) = D(p_1(x)'\beta_1) = A'\beta_1$ gives

$$\begin{aligned} \sqrt{n}F[D(\widehat{g}_1) - D(\bar{g}_1)] &= \sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_1 - \beta_1] \\ &\quad + \sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_2] + \sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon] \end{aligned}$$

The terms $\sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_1 - \beta_1]$ and $\sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_2]$ are negligible while the third term $\sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon]$ satisfies a central limit theorem. First, note the expressions $1_{\mathscr{L}}\|FA'\widehat{Q}^{-1}\| = O_P(1)$, $1_{\mathscr{L}}\|FA'\widehat{Q}^{-1/2}\| = O_P(1)$ both hold by arguments in Newey (1997).

Beginning with the first term,

$$\begin{aligned}
&1_{\mathscr{L}}|\sqrt{n}FA'[\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_1/n - \beta_1]| \\
&= 1_{\mathscr{L}}|\sqrt{n}FA'[(P_1'\mathscr{M}_{\widehat{I}}P_1/n)^{-1}P_1'\mathscr{M}_{\widehat{I}}(G_1 - P_1\beta_1)/n]| \\
&\leqslant 1_{\mathscr{L}}\|FA'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}/\sqrt{n}\|\|G_1 - P_1\beta_1\| \\
&\leqslant 1_{\mathscr{L}}\|FA'\widehat{Q}^{-1/2}\|\sqrt{n}\max_{i\leqslant n}|g_1(x_i) - \bar{g}_1(x_i)| \\
&\leqslant 1_{\mathscr{L}}\|FA'\widehat{Q}^{-1/2}\|\sqrt{n}|g_1 - \bar{g}_1|_0 = O_P(1)O_P(\sqrt{n}K_1^{-\alpha}) = o_P(1)
\end{aligned}$$

Next, consider $\sqrt{n}FA'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_2/n$. By $P_1 = m + W$, triangle inequality, Cauchy-Schwartz and Lemma 4,

$$\begin{aligned}
|FA'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}| &\leqslant |FA'\widehat{Q}^{-1}m'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}| + \|FA'\widehat{Q}^{-1}W'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| \\
&\leqslant \|FA'\widehat{Q}^{-1}\|\|m'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| + \|FA'\widehat{Q}^{-1}\|\|W'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| \\
&= O_P(1)o_P(1) + O_P(1)o_P(1)
\end{aligned}$$

Next consider the last remaining term for which a central limit result will be shown. Note that using the bounds for $\|P_1'\mathscr{M}_{\widehat{I}}\epsilon/\sqrt{n}\|$ and $\|m'\mathscr{M}_{\widehat{I}}\epsilon/\sqrt{n}\|$ derived in Lemma 4,

$$\begin{aligned}
\sqrt{n}FA'\widehat{Q}^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n &= \sqrt{n}FA'Q^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n + \sqrt{n}FA'(\widehat{Q}^{-1} - Q^{-1})P_1'\mathscr{M}_{\widehat{I}}\epsilon/n \\
&= \sqrt{n}FA'Q^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n + O(\|FA'(\widehat{Q}^{-1} - Q^{-1})\|\|P_1'\mathscr{M}_{\widehat{I}}\epsilon/\sqrt{n}\|) \\
&= \sqrt{n}FA'Q^{-1}P_1'\mathscr{M}_{\widehat{I}}\epsilon/n + o_P(1) \\
&= \sqrt{n}FA'Q^{-1}W'\epsilon/n + \sqrt{n}FA'Q^{-1}(W' - P_1'\mathscr{M}_{\widehat{I}})\epsilon/n + o_P(1) \\
&= \sqrt{n}FA'Q^{-1}W'\epsilon/n + O(\|FA'Q^{-1}\|\|m'\mathscr{M}_{\widehat{I}}\epsilon/\sqrt{n}\|) + o_P(1) \\
&= \sqrt{n}FA'Q^{-1}W'\epsilon/n + o_P(1)
\end{aligned}$$

Let $Z_{in} = FA'W_i\epsilon_i/\sqrt{n}$. Then $\sum_i Z_{in} = FA'V'\epsilon/\sqrt{n}$. For each n, Z_{in} is i.i.d. with $E[Z_{in}] = 0, \sum_i E[Z_{in}^2] = 1$. In addition,

$$nE[1_{\{|Z_{in}|>\delta\}}Z_{in}^2] = n\delta^2 E[1_{\{|Z_{in}/\delta|>1\}}Z_{in}^2/\delta^2] \leqslant n\delta^2 E[Z_{in}^4/\delta^4]$$

$$\leqslant n\delta^2\|FA'\|^4\zeta_0(K_1)^2E[\|w_i\|^2E[\epsilon_i^4|x_i]]/n^2\delta^4 \leqslant C\zeta_0(K_1)^2K_1/n \to 0.$$

By the Lindbergh-Feller Central Limit Theorem, $\sum_i Z_{in} \xrightarrow{d} N(0,1)$.

Next consider the plug in variance estimate. First, bound $\|\widehat{A} - A\|$. In the case that $a(g)$ is linear in g, then $a(p_1'\beta) = A'\beta \implies \widehat{A} = A$. Therefore, it is sufficient to consider the case (ii) of Assumption 7, that $a(g)$ is not linear in g. For ν as in the statement of Assumption 7, Define the event $\mathcal{E} = \mathcal{E}_n = \{|\widehat{g}_1 - g_1|_d < \nu/2\}$. In addition, let $\widehat{J} =$

$(D(p_{11};\widehat{g}_1),...,D(p_{1K};\widehat{g}_1))'$. Then for any β such that $|p_1'\beta-\widehat{g}|<\nu/2$, it follows that $|p_1'\beta-g_1|\leq\nu$ and

$$\begin{aligned}1_{\mathcal{E}}|a(p_1'\beta)-a(\widehat{g}_1)-\widehat{J}'(\beta-\widehat{\beta})|/\|\beta-\widehat{\beta}\|&\\ =1_{\mathcal{E}}|a(p_1'\beta-a(\widehat{g})-D(p_1'\beta;\widehat{g})+D(\widehat{g};\widehat{g})|/\|\beta-\widehat{\beta}\|&\\ \leqslant 1_{\mathcal{E}}C\cdot|p_1'\beta-\widehat{g}|_d^2/\beta-\widehat{\beta}\|\leqslant 1_{\mathcal{E}}C\cdot\zeta_d(K_1)^2\|\beta-\widehat{\beta}\|\to 0&\end{aligned}$$

Therefore, $\widehat{A}$ exists and equals $\widehat{J}$ if $1_{\mathcal{E}}=1$.

$$\begin{aligned}1_{\mathcal{E}}\|\widehat{A}-A\|^2=1_{\mathcal{E}}(\widehat{A}-A)'(\widehat{A}-A)=1_{\mathcal{E}}|D((\widehat{A}-A)'p_1;\widehat{g})-D((\widehat{A}-A)'p_1;g_1)|&\\ \leqslant C\cdot 1_{\mathcal{E}}|(\widehat{A}-A)'p_1|_d|\widehat{g}-g_1|_d\leqslant C\cdot\|\widehat{A}-A\|\zeta_d(K_1)|\widehat{g}-g_0|_d&\end{aligned}$$

This gives $1_{\mathcal{E}}\|\widehat{A}-A\|\leqslant C\cdot\zeta_d(K_1)|\widehat{g}-g_1|_d=O_P(\zeta_d(K_1)^2(\sqrt{K}/\sqrt{n}K^{-\alpha}))\xrightarrow{p}0$.

A consequence of the bound on $\|\widehat{A}-A\|$ is that $1_{\mathcal{E}}\|F\widehat{A}\|\leqslant 1_{\mathcal{E}}\|F\|\|\widehat{A}-A\|+\|FA\|=O_P(1)$. Similarly, $1_{\mathcal{E}}\|F\widehat{A}\widehat{Q}^{-1}\|=O_P(1)$. Next, define $\widehat{h}=1_{\mathcal{E}}\widehat{Q}^{-1}\widehat{A}F$ and $h=1_{\mathcal{E}}Q^{-1}AF$.

$$\begin{aligned}\|\widehat{h}-h\|&\leqslant 1_{\mathcal{E}}\|F\widehat{A}'\widehat{Q}^{-1}(Q-\widehat{Q})\|+1_{\mathcal{E}}\|F(\widehat{A}-A)'\|\\ &\leqslant 1_{\mathcal{E}}\|F\widehat{A}'\widehat{Q}^{-1}\|\|Q-\widehat{Q}\|+1_{\mathcal{E}}|F|\|\widehat{A}-A\|\xrightarrow{p}0\end{aligned}$$

Next, note that $h'\Sigma h=1_{\mathcal{E}}$. In addition, $\Sigma\leqslant C\cdot I$ in the positive definite sense by Assumption. Therefore,

$$\begin{aligned}1_{\mathcal{E}}|\widehat{h}'\Sigma\widehat{h}-1|=|\widehat{h}\Sigma\widehat{h}-h'\Sigma h|\leqslant(\widehat{h}-h)'\Sigma(\widehat{h}-h)+|2(\widehat{h}-h)'\Sigma h|&\\ \leqslant C\cdot\|\widehat{h}-h\|^2+2((\widehat{h}-h)'\Sigma(\widehat{h}-h))^{1/2}(h'\sigma h)^{1/2}&\\ \leqslant o_P(1)+C\|\widehat{h}-h\|\xrightarrow{p}0.&\end{aligned}$$

Define $\tilde{\Sigma}=\sum_i W_iW_i'\epsilon_i^2/n$, an infeasible sample analogue of Σ. By reasoning similar to that showing $\|\bar{Q}-Q\|\xrightarrow{p}0$ it follows that $\|\tilde{\Sigma}-\Sigma\|\xrightarrow{p}0$. Then this implies that $1_{\mathcal{E}}|\widehat{h}\tilde{\Sigma}\widehat{h}-\widehat{h}'\Sigma\widehat{h}|=|\widehat{h}'(\tilde{\Sigma}-\Sigma)\widehat{h}|\leqslant\|\widehat{h}\|^2\|\tilde{\Sigma}-\Sigma\|=O_P(1)o_P(1)\xrightarrow{p}0$.

Next, let $\Delta_{1i}=g_1(x_i)-\widehat{g}_1(x_i)$ and $\Delta_{2i}=g_2(x_i)-\widehat{g}_2(x_i)$. Then $\max_{i\leqslant n}|\Delta_i|\leqslant|\widehat{g}_1-g_1|_0=o_(1)\xrightarrow{p}0$ follows from the proposition above. Let $\tilde{S}=\sum_i W_iW_i'|\epsilon_i|/n$, $\widehat{S}=\sum_i\widehat{W}_i\widehat{W}_i'|\epsilon_i|/n$ and $S=E[W_iW_i'|\epsilon_i|]=E[W_iW_i'E[|\epsilon_i||x_i]]\leqslant C\cdot Q$. Let $\omega_i^2=\widehat{h}'W_iW_i'\widehat{h}$. Bound $\tilde{\Sigma}$ to $\widehat{\Sigma}$ by considering the quantity

$$\begin{aligned}\mathcal{E}_n|F\widehat{V}F-\widehat{h}'\tilde{\Sigma}\widehat{h}|=|\widehat{h}'(\widehat{\Sigma}-\tilde{\Sigma})\widehat{h}|=\left|\sum_{i=1}^n\widehat{h}'\widehat{W}_i\widehat{W}_i'\widehat{\epsilon}_i^2\widehat{h}/n-\sum_{i=1}^n\widehat{h}'W_iW_i'\epsilon_i^2\widehat{h}/n\right|&\\ \leqslant\left|\sum_{i=1}^n\omega_i^2(\widehat{\epsilon}_i^2-\epsilon_i^2)/n\right|+\left|\sum_{i=1}^n(\widehat{\omega}_i^2-\omega_i^2)\epsilon_i^2/n\right|&\end{aligned}$$

Both terms on the right hand side will be bounded. Consider the first term. Expanding $(\widehat{\epsilon}_i^2 - \epsilon_i^2)$ gives

$$\left|\sum_{i=1}^n \omega_i^2(\widehat{\epsilon}_i^2 - \epsilon_i^2)/n\right| \leqslant 2\left|\sum_{i=1}^n \omega_i^2\Delta_{1i}^2/n\right| + 2\left|\sum_{i=1}^n \omega_i^2\Delta_{2i}^2/n\right| + 2\left|\sum_{i=1}^n \omega_i^2\Delta_{1i}\epsilon_i/n\right| + 2\left|\sum_{i=1}^n \omega_i^2\Delta_{2i}\epsilon_i/n\right|$$

These four terms above are bounded in order of their appearence.

$$\sum_{i=1}^n \omega_i^2\Delta_{1i}^2/n \leqslant \max_{i\leqslant n}|\Delta_{1i}|\sum_{i=1}^n \omega_i^2/n = o_P(1)O_P(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{2i}^2/n \leqslant \max_{i\leqslant n}|\Delta_{2i}|\sum_{i=1}^n \omega_i^2|\epsilon_i|/n = o_P(1)O_P(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{1i}\epsilon_i/n \leqslant \max_{i\leqslant n}|\Delta_{1i}|\sum_{i=1}^n \omega_i^2|\epsilon_i|/n = o_P(1)O_P(1)$$
$$\sum_{i=1}^n \omega_i^2\Delta_{2i}\epsilon_i/n \leqslant \max_{i\leqslant n}|\Delta_{2i}|\sum_{i=1}^n \omega_i^2|\epsilon_i|/n = o_P(1)O_P(1)$$

Where the bounds $\max_{i\leqslant n}|\Delta_{1i}| = o_P(1)$ follows by the proposition and $\max_{i\leqslant n}|\Delta_{2i}| = o_P(1)$ follows from Lemma 5 below. On the other hand, the second term is bounded by

$$\begin{aligned}
&\left|\sum_{i=1}^n \widehat{h}(\widehat{W}_i\widehat{W}_i' - W_iW_i')\epsilon_i^2\widehat{h}/n\right| \leqslant \max_{i\leqslant n}|\epsilon_i^2|\left|\sum_{i=1}^n \widehat{h}(\widehat{W}_i\widehat{W}_i' - W_iW_i')\widehat{h}/n\right| \\
&\leqslant \max_{i\leqslant n}|\epsilon_i^2|\|\widehat{h}\|^2\|\sum_{i=1}^n(\widehat{W}_i\widehat{W}_i' - W_iW_i')/n\| = \max_{i\leqslant n}|\epsilon_i^2|\|\widehat{h}\|^2\|\widehat{Q} - \bar{Q}\| \\
&= O_P(n^{2/q})O_P(1)O_P(K_1^{3/2}\phi/\sqrt{n}) \\
&= o_P(1)
\end{aligned}$$

by Assumption 4. This implies that $\mathcal{E}_n|F\widehat{V}F - 1| \xrightarrow{p} 0$. With probability approaching 1, $1_{\mathcal{E}} = 1$, this gives $F\widehat{V}F \xrightarrow{p} 1$ which in turn implies that

$$\sqrt{n}\widehat{V}^{-1/2}(\widehat{\theta} - \theta) = \sqrt{n}F(\widehat{\theta} - \theta)/(F\widehat{V}F)^{1/2} \xrightarrow{d} N(0,1).$$

To provide a rate of convergence, $|V| \leqslant C\cdot\zeta_d(K_1)^2$ since $\widehat{\theta} = \theta_0 + (V^{1/2}/\sqrt{n})\sqrt{n}F(\widehat{\theta} - \theta) = \theta + O_P(V^{1/2}/\sqrt{n})$. Cauchy-Schwartz inequality implies that $|p_1'\beta|_d \leqslant \zeta_d(K_1)\|\beta\|$ for any choice of β. Then $\|A\|^2 = |D(p_1'A)| \leqslant C\cdot|p_1'A|_d \leqslant C\cdot\zeta_d(K_1)\|A\|$. This gives $\|A\| \leqslant C\cdot\zeta_d(K_1)$ and $|V| \leqslant C\cdot\|A\|^2 \leqslant C\cdot\zeta_d(K_1)^2$.

The proof of the second statement of the Theorem uses similar arguments as the proof of the first and follows from the proof of Theorem 3 in Newey (1997).

9.4. **Lemmas.** The first lemma is a performance bound for Post-Lasso estimates. It is required for use in the next two Lemmas. It is based on the results of Belloni, Chen, Chernozhukov and Hansen (2013). Define the following four events which are useful for describing the regularization properties of the lasso regressions.

$$\mathcal{A}_{FS} = \{\lambda^{FS}/n \geqslant c\|S_k\|_\infty \forall k\}, \quad \mathcal{A}_{RF} = \{\lambda^{RF}/n \geqslant c\|S\|_\infty\}$$

$$\mathcal{B}_{FS} = \{\ell\Psi_{jk}^{FS} \leqslant \widehat{\Psi}_{jk}^{FS} \leqslant u\Psi_{ij}^{FS} \forall j,k\}, \quad \mathcal{B}_{RF} = \{\ell\Psi_{j}^{RF} \leqslant \widehat{\Psi}_{j}^{RF} \leqslant u\Psi_{j}^{RF} \forall j\}.$$

Define the regularization event $\mathscr{R} = \mathcal{A}_{FS} \cap \mathcal{A}_{RF} \cap \mathcal{B}_{FS} \cap \mathcal{B}_{RF}$. In addition, define $c_0 = (uc+1)/(\ell c - 1)$. Let $\kappa_C = \min_{\delta \in \Delta_{C,T}|T|\leqslant s} s\delta'(P_2'P_2/n)\delta/\|\delta_T\|_1^2$ where $\Delta_{C,T} = \{\delta \neq 0 : \|\delta_{T^c}\|_1 \leqslant C\|\delta_T\|_1\}$. This defines the restricted eigenvalue and is useful for Lasso bounds. For more details regarding the definition, see for example, Bickel, Ritov, and Tybakov (2009). Let $\kappa_{c_0}^k = \min_{\|\Psi_k^{FS}\delta_{T_k^c}\|_1 \leqslant c_0\|\Psi_k^{FS}\delta_{T_k}\|_1\|} \sqrt{s}\delta'(P_2'P_2/n)\delta/\|\widehat{\Psi}_k^{FS}\delta_{T_k}\|_1$. Define κ_{c_0} analogously using the reduced form $\widehat{\Psi}^{RF}$.

Lemma 1. *Under the conditions given in Assumption 4, the following inequalities holds.*

$$1_{\mathscr{R}} \max_{k\leqslant K_1} \|\mathscr{M}_{\widehat{I}} m_k/\sqrt{n}\|_2 \leqslant 1_{\mathscr{R}} \left(\max_{k\leqslant K_1} (u+1/c) \frac{\lambda^{FS}\sqrt{s}}{n\kappa_{c_0}^l} + 3\xi FS \right).$$

$$1_{\mathscr{R}} \|\mathscr{M}_{\widehat{I}} E[y|\mathcal{G}_2(x)]/\sqrt{n}\|_2 \leqslant 1_{\mathscr{R}} \left((u+1/c) \frac{\lambda^{RF}\sqrt{s}}{n\kappa_{c_0}} + 3\xi_{RF} \right).$$

In addition, the regularization event satisfies $\mathrm{P}(\mathscr{R}) \to 1$.

Proof. That $\mathcal{A}_{RF}$ holds with probability approaching 1 was established in Belloni, Chen, Chernozhukov and Hansen (2013). The conditions listed in Assumption 6 allow use of the same argument to show that $\mathcal{A}_{FS}$ holds with high probability by allowing the application of the moderate deviation results of de la Pena, Lai and Shao (2009). The proof of that fact is an identical to the proof given in BCCH and omitted. In addition, $\mathcal{B}_{RF}$ holds with probability approaching one by Lemma 11 of Belloni, Chen, Chernozhukov and Hansen (2013). Again, under the additional conditions listed in Assumption 6, the argument extends to show $\mathcal{B}_{FS}$ happens with probability approaching 1.

Define $\bar{p}_{1k}(x_i) = p_{1k}(x_i) - \sum_{i=1}^n p_{1k}(x_i)$ and $\tilde{p}_{1k}(x_i) = p_{1k}(x_i) - E[p_{1k}(x_i)]$. Then let

$$\left(\widehat{\Psi}_{jk}^{FS}\right)^2 = \sum_{i=1}^n p_{2j}(x_i)^2 \bar{p}_{1k}(x_i)^2/n, \;\; \left(\tilde{\Psi}_{jk}^{FS}\right)^2 = \sum_{i=1}^n p_{2j}(x_i)^2 \tilde{p}_{1k}(x_i)^2/n$$

To show $\mathrm{P}(\mathcal{B}_{FS}) \to 1$ for the basic penalty loadings it is sufficient to show that $u_1 := \max_{k,j} |(\widehat{\Psi}_{jk}^{FS})^2 - (\tilde{\Psi}_{jk}^{FS})^2| \xrightarrow{p} 0$ and $u_2 := \max_{k,j} |(\widehat{\Psi}_{jk}^{FS})^2 - (\Psi_{jk}^{FS})^2| \xrightarrow{p} 0$. Assumption six

gives $u_2 \xrightarrow{p} 0$. Next note that that using $\sum_{i=1}^n \tilde{p}_{1k}(x_i)/n - E[p_{1k}(x_i)]$,

$$\begin{aligned} u_1 =& \max_{k,j} |\sum_{i=1}^n p_{2j}(x_i)^2 \left[(\tilde{p}_{1k}(x_i) - \sum_{i=1}^n \tilde{p}_{1k}(x_i)/n)^2 - \tilde{p}_{1k}(x_i)^2 \right]| \\ \leqslant & \max_{k,j} 2|(\sum_{i=1}^n p_{2j}(x_i)^2 \tilde{p}_{1k}(x_i)/n)| \sum_{i=1}^n \tilde{p}_{1k}(x_i)/n)| \\ & + \max_{k,j} (\sum_{i=1}^n p_{2j}(x_i)^2)(\sum_{i=1}^n \tilde{p}_{1k}(x_i)/n)^2 \end{aligned}$$

Note that $\max_{k,j}(\sum_{i=1}^n p_{2j}(x_i)\tilde{p}_{1k}(x_i)^2/n) \leqslant \max_{i,k,j} |p_{2j}(x_i)| \max_{k,j} \sqrt{\sum_{i=1}^n p_{2j}(x_i)^2 \tilde{p}_{1k}(x_i)^2/n}$ and by Assumption 6. The second term converges to zero by Assumption 6.

The proof that the iterated option mentioned in the text is valid follows similar logic. In addition, it is shown in BCCH for fixed K_1. The result follows from their proof, but using Lemma 3 of this paper for a bound on $\max_k \|m_k' \mathscr{M}_{\widehat{I}}\|$.

Therefore, $1_{\mathscr{R}} \xrightarrow{p} 1$ giving the last claim of the lemma. The first two claims follow immediately from the third statement of Lemma 7 in Belloni, Chen, Chernozhukov and Hansen (2013).

□

Lemma 2. $\max_{k \leqslant K_1}(1/\kappa_{c_0}^k) = O_P(1)$ *and* $\max_{k \leqslant K_1} |\{j : \widehat{\Gamma}_{kj} \neq 0\}| = O_P(s)$

Proof. For the first result, let $a = \min_{k,j} |\widehat{\Psi}_{kj}^{FS}|_\infty$ and $b = \max_{k,j} |\widehat{\Psi}_{kj}^{FS}|_\infty$. Step 1 of the the proof of Theorem 1 in BCCH shows that $\max_{k \leqslant K_1}(1/\kappa_{c_0}^k) \leqslant b(\kappa_{bc_0/a}(P_2'P_2/n))^{-1}$. By the results of the previous lemma, a and b are bounded from above and away from zero with probability approaching 1. Then using Assumption 5, $(\kappa_{bc_0/a}(P_2'P_2/n))^{-1} = O_P(1)$. This implies the first statement. For the second statement, let $\widehat{s}_k$ be the number of incorrectly selected terms in the k-th first stage regression. Then By Lemma 10 of BCCH, $\widehat{s}_k \leqslant s\varphi_{\max}(q) \max_j |\widehat{\Psi}_{jk}^0|^4 (2c_0/\kappa_{c_0}^k + 6c_0 n\xi_F S/\lambda\sqrt{s})^2$ for every integer $q > 2s\varphi_{\max}(q) \max_j |\widehat{\Psi}_{jk}^0|^{-2}(2c_0/\kappa_{c_0}^k + 6c_0 n\xi_F S/\lambda\sqrt{s})^2$. The choice $q = \kappa'' 2s\varphi_{\max}(q) \max_{k,j} |\widehat{\Psi}_{jk}^0|^{-2}(2c_0/\kappa_{c_0}^k + 6c_0 n\xi_F S/\lambda\sqrt{s})^2$ yields $\max_k \widehat{s}_k \leqslant$ gives $\varphi_{\max}(q) = O_P(1)$ by Assumption 5 and by using $\max_{k,j} |\widehat{\Psi}_{jk}^0|^{-2} = O_P(1)$, $\max_k 2c_0/\kappa_{c_0}^k = O_P(1)$ and $6c_0/n\xi_{FS}/\lambda^{FS}\sqrt{s} = o_P(1)$ which were shown in Lemma 1. □

The following lemmas bounds various quantities used in the proof above. The lemma provides analogous results to steps 4-6 of Belloni, Chernozhukov and Hansen (2013)'s proof of their Theorem 1 but accounts increasing number of series terms terms in the first dictionary.

Lemma 3. *First Stage and Reduced Form Performance Bounds*

(i) $\max_{k\leqslant K_1}\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\| = O_P(\sqrt{\phi/n}+\xi_{FS}),\quad \xi_{FS}=O_P(\phi/n)$

(ii) $\|\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| = O_P(\sqrt{K_1\phi/n}+\xi_{RF}),\quad \xi_{RF}=O(K_1^{-\alpha})+O_P(\sqrt{K_1\phi/n})$

(iii) $\max_{k\leqslant K_1}\|\widehat{\Gamma}_k(\widehat{I})-\Gamma_k\| = O_P(\sqrt{\phi/n})$

(iv) $\|b(G_2;\widehat{I})-\beta_2\| = O_P(\sqrt{\phi/n}+K_1^{-\alpha})$

(v) $\max_{k\leqslant K_1}\|b(W_k;\widehat{I}\|_1 = O_P(\sqrt{s\phi/n})$

(vi) $\max_{k\leqslant K_1}\|P_2'W_k/\sqrt{n}\|_\infty = O_P(\phi/s), \|P_2'\epsilon/\sqrt{n}\|_\infty = O_P(K_1\sqrt{s\phi/n})$.

Proof. Statement (i) follows from an application of Lemma 1:

$$\begin{aligned}
1_{\mathscr{R}}\max_{k\leqslant K_1}\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\| &\leqslant 1_{\mathscr{R}}\max_{k\leqslant K_1}\|\mathscr{M}_{\widehat{I}_k}m_k/\sqrt{n}\| \\
&\leqslant 1_{\mathscr{R}}\max_{k\leqslant K_1}(u+1/c)\lambda^{FS}\sqrt{s}/n\kappa_{c0}^k + 3\xi_{FS} \\
&\leqslant 1_{\mathscr{R}}\max_{k\leqslant K_1}(u+1/c)(C\sqrt{n\log(\max(K_1K_2,n))}\sqrt{s}/n\kappa_{c_0}^k + 1_{\mathscr{R}}3\xi_{FS} \\
&= O(\sqrt{\phi/n})/\kappa_{c_0}^k + 3\xi_{FS} = O_P(\sqrt{\phi/n})
\end{aligned}$$

Where the last equality follows from Lemma 2 and the definition of λ^{FS}, $\xi_{FS}=O_P(\sqrt{s/n})$ and $1_{\mathscr{R}}\xrightarrow{p}1$. Next, Consider statement (ii).

$$\begin{aligned}
1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| &\leqslant 1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}(E[G_1|\mathcal{G}_2]+G_2)/\sqrt{n}\| + 1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}E[G_1|\mathcal{G}_2]/\sqrt{n}\| \\
&\leqslant 1_{\mathscr{R}}\|\mathscr{M}_{I_{RF}}(E[G_1|\mathcal{G}_2]+G_2)/\sqrt{n}\| + 1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}E[G_1|\mathcal{G}_2]/\sqrt{n}\| \\
&\leqslant 1_{\mathscr{R}}(u+1/c)\lambda^{RF}\sqrt{K_1s}/n\kappa_{c_0}^{RF} + 1_{\mathscr{R}}3\xi_{RF} + 1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}E[G_1|\mathcal{G}_2]/\sqrt{n}\| \\
&\leqslant 1_{\mathscr{R}}(u+1/c)(C\sqrt{n\log(\max(K_1K_2,n))}\sqrt{K_1s}/n\kappa_{c_0}^{RF} \\
&\quad + 1_{\mathscr{R}}3\xi_{RF} + 1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}E[G_1|\mathcal{G}_2]/\sqrt{n}\|
\end{aligned}$$

The Last bound follows from Lemma 1. To control the approximation error for the reduced form, ie. to bound ξ_{RF}, note that

$$g_1(x)+g_2(x) = p_2(x)'(\Gamma\beta_1+\beta_2) + (g_1(x)-p_1(x)\beta_1) + (g_2(x)-p_2(x)'\beta_2)$$

The approximation error ξ_{RF} is then given by

$$\begin{aligned}
\xi_{RF}^2 &= (G_1-P_1\beta_1+G_2-P_2\beta_2)'(G_1-P_1\beta_1+G_2-P_2\beta_2)/n \\
&\leqslant 2(G_1-P_1\beta_1)'(G_1-P_1\beta_1)/n + 2(G_2-P_2\beta_2)'(G_2-P_2\beta_2)/n \\
&= O(K^{-2\alpha}) + O(K_1s/n)
\end{aligned}$$

Next consider $\|\mathscr{M}_{\widehat{I}}E[G_1|\mathcal{G}_2]/\sqrt{n}\|$ and note that $E[G_1|\mathcal{G}_2] = m\beta_1 + E[G_1-m\beta_1|\mathcal{G}_2]$. By statement (i) of this lemma, $\|\mathscr{M}_{\widehat{I}}m\beta_1/\sqrt{n}\| \leqslant \max_k\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\|\|\beta_1\| = O_P(\sqrt{\phi/n})O(1)$.

Next, $\|\mathscr{M}_{\widehat{I}}E[G_1 - m\beta_1|\mathcal{G}_2]/\sqrt{n}\| \leqslant \|\mathscr{M}_{\widehat{I}}E[G_1 - P_1\beta_1|\mathcal{G}_2]/\sqrt{n}\| + \|\mathscr{M}_{\widehat{I}}E[P_1\beta_1 - m\beta_1|\mathcal{G}_2]/\sqrt{n}\|$. The first term $\|\mathscr{M}_{\widehat{I}}E[G_1 - P_1\beta_1|\mathcal{G}_2]/\sqrt{n}\|$ is $O(K^{-\alpha})$ and the second term $\|\mathscr{M}_{\widehat{I}}E[P_1\beta_1 - m\beta_1|\mathcal{G}_2]/\sqrt{n}\|$ vanishes identically. These results put together establish that $1_{\mathscr{R}}\|\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| \leqslant 1_{\mathscr{R}}O_P(\sqrt{K_1\phi/n}) + O(K^{-\alpha})$. The result follows by noting that $1_{\mathscr{R}} \xrightarrow{p} 1$.

Next consider statement (iii). Let $\widehat{T} = \widehat{I} \cup \text{ supp}(\Gamma_1) \cup ... \cup \text{ supp}(\Gamma_{K_1})$.

$$\max_{k\leqslant K_1} \|\widehat{\Gamma}_k(\widehat{I}) - \Gamma_k\| \leqslant \max_k \{\sqrt{\varphi_{\min}(|\widehat{T}_k|)}\|\widehat{\Gamma}_k(\widehat{I}) - \Gamma_k\|\} \leqslant \max_{k\leqslant K_1} \|P_2(\widehat{\Gamma}_k(\widehat{I}) - \Gamma_k)/\sqrt{n}\|$$
$$\leqslant \max_{k\leqslant K_1}\{\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\| + \|(m_k - P_2\Gamma_k)/\sqrt{n}\|\} = O_P(\sqrt{\phi/n}).$$

Where the last bound follows from $\varphi_{\min}(\widehat{T}) = O_P(1)$ by Assumption 5 on the restricted eigenvalues and by $\widehat{T} = O_P(K_1 s)$ by the result of the lemma above. Statement (iv) follows from similar reasoning as for statement (iii).

To show statement (v), note that by Lemma 4 of BCH, a sufficient condition for

$$\max_{k\leqslant K_1 j\leqslant K_2} \frac{|P_{2j}'W_k|/\sqrt{n}}{\sqrt{\sum_i p_{2j}(x_i)^2 W_{ki}^2/n}} = O_P(\phi/s)$$

Is that $\min_{k\leqslant K_1, j\leqslant K_2} E[p_{2j}(x_i)^2 W_{ki}^2]^{1/2} E[|p_{2j}(x_i)|^3|W_{ki}|^3]^{-1/3} = O(1)$ and $\log(K_1K_2) = o(n^{1/3})$. These conditions follow from Assumption 6. In addition, $\sqrt{\sum_i p_{2j}(x_i)^2 W_{ki}^2/n} = O_P(1)$ by Assumption 6. This gives the first part of statement (vi). The second part follows in the same manner.

Statement (v):

$$\max_{k\leqslant K_1} \|b(W_k;\widehat{I})\|_1 \leqslant \max_{k\leqslant K_1} \sqrt{|\widehat{I}|}\|b(W_k;\widehat{I})\| \leqslant \max_{k\leqslant K_1} \sqrt{|\widehat{I}|}(P_2(\widehat{I})'P_2(\widehat{I}))^{-1}P_2(\widehat{I})'W_k/n\|$$
$$\leqslant \sqrt{|\widehat{I}|}\varphi_{\min}^{-1}(|\widehat{I}|)\sqrt{|\widehat{I}|}\max_{k\leqslant K_1}\|P_2'W_k/\sqrt{n}\|_\infty = O_P(K_1\sqrt{s}\sqrt{\phi/n})$$

□

Lemma 4. *The following bounds hold.*

(i) $\|W'\mathscr{P}_{\widehat{I}}W/n\| = O_P(K_1\phi/n)$

(ii) $\|m'\mathscr{M}_{\widehat{I}}m/n\| = O_P(K_1\phi/n)$

(iii) $\|m'\mathscr{M}_{\widehat{I}}W/n\| = O_P(K_1^{3/2}\sqrt{\phi^2/n})$

(iv) $\|m'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| = O_P(\sqrt{K_1\phi/n}(\sqrt{n}K_1^{-\alpha} + \sqrt{K_1^2\phi^2/n}))$

(v) $\|W'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| = O_P(\zeta_0(K_1)\sqrt{K_1\phi/n} + \sqrt{K_1\phi^2/n} + K_1^{-\alpha}\sqrt{K_1\phi/s})$

(vi) $\|W'\mathscr{P}_{\widehat{I}}\epsilon/\sqrt{n}\| = O_P(K_1^{3/2}\sqrt{\phi^2/n})$

(vii) $\|m'\mathscr{M}_{\widehat{I}}\epsilon/\sqrt{n}\| = O_P(K_1\sqrt{\phi/n})$

Proof. Bounds for statement (i):

$$\begin{aligned}\|W'\mathscr{P}_{\widehat{I}}W/n\|^2 &= \sum_{k,l\leqslant K_1}(W_k'\mathscr{P}_{\widehat{I}}W_l/n)^2 = \\ &= \sum_{k,l\leqslant K_1}(b(W_k;\widehat{I})'P_2W_l/n)^2 \leqslant \sum_{k,l\leqslant K_1}\|b(W_k;\widehat{I})/\sqrt{n}\|_1^2\|P_2'W_l/\sqrt{n}\|_\infty^2 \\ &= \left(\sum_{k\leqslant K_1}\|b(W_k;\widehat{I}_k)\|_1^2/n\right)\left(\sum_{k\leqslant K_1}\|P_2'W_l/\sqrt{n}\|_\infty^2\right) \\ &= O_P(K_1s\phi/n^2)O_P(K_1\phi/s)\end{aligned}$$

Where the last probability bounds follow from Lemma 3. This gives $\|W'\mathscr{P}_{\widehat{I}}W/n\| = o_P(1)$ by $K_1^2\phi^2/n \to 0$.

Next, bounds for statement (ii):

$$\begin{aligned}\|m'\mathscr{M}_{\widehat{I}}m/n\|^2 &= \sum_{k,l\leqslant K_1}(m_k'\mathscr{M}_{\widehat{I}}m_l/n)^2 \leqslant \sum_{k,l\leqslant K_1}\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\|^2\|\mathscr{M}_{\widehat{I}}m_l/\sqrt{n}\|^2 \\ &= \left(\sum_{k\leqslant K_1}\|\mathscr{M}_{\widehat{I}}m_k/\sqrt{n}\|^2\right)^2 = O_P(K_1\phi/n)^2\end{aligned}$$

where again the final probability bounds follow from Lemma 3. This implies that $\|m'\mathscr{M}_{\widehat{I}}m/n\| = o_P(1)$ by $K_1\phi/n \to 0$. Finally, a bound on the third term is established by

$$\begin{aligned}\|m'\mathscr{M}_{\widehat{I}}W/n\| &= \|m'W/n - m'\mathscr{P}_{\widehat{I}}W/n\| \\ &= \|\Gamma P_2'W/n + (m' - \Gamma'P_2')W/n - m'\mathscr{P}_{\widehat{I}}W/n\| \\ &= \|R_m'W/n + (\Gamma - \Gamma(\widehat{I}))'P_2'W/n\| \\ &\leqslant \|R_m'W/n\| + \|(\Gamma - \Gamma(\widehat{I}))'P_2'W/n\|\end{aligned}$$

Then the first term in the last line is bounded by $\|R_m'W/n\| \leqslant \|(m'-\Gamma'P_2')(m-P_2\Gamma)/n\|2\zeta_0(K_1)$ while the second term has

$$\begin{aligned}
\|(\Gamma - \Gamma(\widehat{I}))'P_2'W_l/n\|^2 &= \sum_{k,l}[(\Gamma_k - \Gamma_k(\widehat{I}))'P_2'W_l/n]^2 \\
&= \sum_{k,l} \|\Gamma_k - \Gamma_k(\widehat{I})\|_1^2 \|P_2'W_l/\sqrt{n}\|_\infty^2 \\
&= \left(\sum_k \|\Gamma_k - \Gamma_k(\widehat{I})\|_1^2\right)\left(\sum_l \|P_2'W_l/\sqrt{n}\|_\infty^2\right) \\
&\leqslant \left(|\widehat{I}|\sum_k \|\Gamma_k - \Gamma_k(\widehat{I})\|^2\right)\left(\sum_l \|P_2'W_l/\sqrt{n}\|_\infty^2\right) \\
&= O_P(K_1)O_P(K_1 s\phi/n)K_1 O_P(\phi/s)
\end{aligned}$$

With the last asertion following from Lemma 3. this gives $\|m'\mathscr{M}_{\widehat{I}}W/n\| = o_P(1)$

Statement (iv):

$$\begin{aligned}
\|FA'\widehat{Q}^{-1}m'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| &\leqslant \|FA'\widehat{Q}^{-1}\| \max_{k\leqslant K_1} \|m'\mathscr{M}_{\widehat{I}}/\sqrt{n}\|\sqrt{n}\|\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| \\
&= O_P(1)O_P(\sqrt{\phi/n})\sqrt{n}O_P(\sqrt{\phi/n} + K^{-\alpha}) = o_P(1).
\end{aligned}$$

Statement (v):

$$\begin{aligned}
\|W'\mathscr{M}_{\widehat{I}}G_2/\sqrt{n}\| &\leqslant \|(G_2 - P_2'\beta_2)'W/\sqrt{n}\| + \|(b(G_2;\widehat{I}) - \beta_2)'P_2'W/\sqrt{n}\| \\
&\leqslant O_P(\zeta_0\sqrt{K_1\phi/n}) + \sqrt{K_1}\max_{k\leqslant K_1}\|b(G_2;\widehat{I}) - \beta_2\|_1\|P_2'W_k/\sqrt{n}\|_\infty \\
&\leqslant O_P(\zeta_0\sqrt{K_1\phi/n}) + \sqrt{K_1}O_P(\sqrt{\phi/n} + K_1^{alpha})O_P(\sqrt{\phi}/s)
\end{aligned}$$

Statement (vi):

$$\begin{aligned}
\|W'\mathscr{P}_{\widehat{I}}\epsilon/\sqrt{n}\| &= \|b(W;\widehat{I})P_2'\epsilon/\sqrt{n}\| \\
&\leqslant \sqrt{K_1}\max_k \|b(W_k;\widehat{I})\|_1\|P_2'\epsilon/\sqrt{n}\|_\infty \\
&= \sqrt{K_1}O_P(\sqrt{K_1^2 s\phi/n})O_P(\sqrt{\phi}/s)
\end{aligned}$$

Statement (vii): Let $R_m = m - P_2\Gamma$. By reasoning similar to that for Lemma 3(iii), $m'\mathscr{M}_{\widehat{I}}\epsilon$

$$\begin{aligned}\|\Gamma(\widehat{I})-\Gamma)'P_2'\epsilon/\sqrt{n}\| &\leqslant \sqrt{K_1}\max_{k\leqslant K_1}|\Gamma(\widehat{I})-\Gamma)'P_2'\epsilon/\sqrt{n}| \\ &\leqslant \sqrt{K_1}\max_{k\leqslant K_1}\|\Gamma(\widehat{I})-\Gamma)\|_1\|P_2'\epsilon/\sqrt{n}\|_\infty \\ &\leqslant \sqrt{K_1}\max_{k}\sqrt{|\widehat{I}|+K_1 s}\|\Gamma(\widehat{I})-\Gamma)\|\|P_2'\epsilon/\sqrt{n}\|_\infty \\ &= O_P(K_1\sqrt{\phi/n})O_P(\sqrt{\phi/s})\end{aligned}$$

□

Lemma 5. $\max_{i\leqslant n}|g_2(x_i)-\widehat{g}_2(x_i)| = o_P(1)$

Proof. Let $\widehat{T}=\widehat{I}\cup \text{supp}(\beta_2)$. Then $\max_i|g_2(x_i)-\widehat{g}_2(x_i)| \leqslant \max_i|g_2(x_i)-p_2(x_i)'\beta_2| + \max_i|\widehat{g}_2(x_i)-p_2(x_i)'\beta_2|$. The first term has the bound $\max_i|g_2(x_i)-p_2(x_i)'\beta_2| = O_P(\sqrt{\phi/n})$ by assumption. A bound on the second term is obtained by the following:

$$\begin{aligned}\max_i|\widehat{g}_2(x_i)-p_2(x_i)'\beta_2|^2 &= \max_i|p_2(x_i)'(\widehat{\beta}_2-\beta_2)|^2 \\ &\leqslant \max_i\|p_{2,\widehat{T}}(x_i)\|^2\|\widehat{\beta}_2-\beta_2\|^2 \\ &\leqslant |\widehat{T}|\max_i\max_{j\leqslant K_2}|p_{2j}(x_i)|^2\|\widehat{\beta}_2-\beta_2\|^2 \\ &\leqslant O_P(K_1 s)\max_i\max_{j\leqslant K_2}|p_{2j}(x_i)|^2\|\widehat{\beta}_2-\beta_2\|^2\end{aligned}$$

Then

$$\|\widehat{\beta}_2-\beta_2\| = \|b(y-\widehat{G}_1;\widehat{I})-\beta_2\| = \|b(G_1;\widehat{I})+b(G_2;\widehat{I})+b(\epsilon;\widehat{I})-b(\widehat{G}_1;\widehat{I})-\beta_2\|$$

$$\leqslant \|b(G_1;\widehat{I})-\beta_2\|+\|b(\epsilon;\widehat{I})\|+\|b(G_1-\widehat{G}_1;\widehat{I})\|$$

First note that $\|b(G_2;\widehat{I})-\beta_2\| = O_P(\sqrt{\phi/n}+K_1^{-1\alpha})$ by Lemma 3. Next, $\|b(\epsilon;\widehat{I})\| \leqslant \sqrt{|\widehat{I}|}\phi_{\min}(\widehat{I})\|P_2'\epsilon/n\|_\infty = O_P(\sqrt{K_1 s})O_P(1)\|P_2'\epsilon/\sqrt{n}\|_\infty/\sqrt{n} = O_P(\sqrt{K_1\phi/n})$. Finally,

$\|b(\widehat{G}_1-G_1;\widehat{I})\| \leqslant \|b(P_1(\beta_1-\widehat{\beta}_1);\widehat{I})\|+\|b(G_1-P_1\beta_1;\widehat{I})\|$. The right term is $\|b(G_1-P_1\beta_1;\widehat{I})\| = O_P(K_1^{-\alpha})$. The left term is bounded by

$$\begin{aligned}\|b(P_1(\beta_1 - \widehat{\beta}_1); \widehat{I})\| &\leqslant \varphi_{\min}(|\widehat{T}|)^{-1}\sqrt{\widehat{T}} \max_j |\sum_i p_{2j}(x_i) p_1(x_i)'(\beta_1 - \widehat{\beta}_1)/n| \\ &\leqslant \varphi_{\min}(|\widehat{T}|)^{-1}\sqrt{\widehat{T}} \max_j \sum_i |p_{2j}(x_i)\|p_1(x_i)/n\|\|(\beta_1 - \widehat{\beta}_1)\| \\ &= O_P(1) O_P(\sqrt{K_1 s})\zeta_0(K_1) O_P(\sqrt{K}_1/n + K_1^{-\alpha}) \max_j \sum_i |p_{2j}(x_i)|/n\end{aligned}$$

These together with Assumption 6 imply that $\max_i |\widehat{g}_2(x_i) - p_2(x_i)'\beta_2| = o_P(1)$ □

FIGURE 1. Simulation Table 1

Table 1. Simulation Results: Low Dimensional Design, Average Derivative

	N = 500			N = 800		
	MAD	Med. Bias	RP 5%	MAD	Med. Bias	RP 5%
	A. High First Stage Signal/Noise, High Structural Signal/Noise					
Post-Double	0.089	0.032	0.060	0.085	0.006	0.052
Post-Single I	0.162	0.159	0.196	0.112	0.091	0.176
Post-Single II	0.645	0.645	0.948	0.500	-0.500	0.800
Series I	0.158	0.152	0.232	0.136	0.136	0.256
Series II	0.097	0.023	0.208	0.090	-0.010	0.092
Infeasible	0.051	0.036	0.120	0.055	0.000	0.060
	B. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.081	0.049	0.080	0.069	-0.027	0.048
Post-Single I	0.116	0.111	0.192	0.074	0.025	0.072
Post-Single II	0.523	0.523	0.868	0.135	-0.098	0.152
Series I	0.124	0.117	0.228	0.089	0.078	0.148
Series II	0.087	0.024	0.236	0.077	-0.029	0.112
Infeasible	0.052	0.042	0.176	0.055	-0.019	0.084
	C. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.137	0.058	0.052	0.113	-0.009	0.064
Post-Single I	0.221	0.211	0.176	0.122	0.096	0.092
Post-Single II	0.665	0.665	0.936	0.484	-0.484	0.660
Series I	0.212	0.207	0.164	0.147	0.138	0.168
Series II	0.153	0.043	0.184	0.123	-0.025	0.072
Infeasible	0.077	0.066	0.144	0.089	-0.007	0.048
	D. Low First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.104	-0.026	0.072	0.103	0.017	0.048
Post-Single I	0.107	0.047	0.076	0.108	0.079	0.088
Post-Single II	0.438	0.438	0.792	0.129	-0.069	0.120
Series I	0.107	0.062	0.084	0.131	0.117	0.156
Series II	0.139	-0.045	0.296	0.112	0.009	0.096
Infeasible	0.058	-0.014	0.088	0.073	0.013	0.056

Note: Results are based on 250 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for six different estimators of the average derivative: the Post-Double proposed in this paper; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only, a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); an estimator that uses a small number of series terms (Series I); an estimator that uses a large number of series terms (Series II); and infeasible estimator that is explicitly given the control function.

FIGURE 2. Simulation Table 2

Table 2. Simulation Results: Low Dimensional Design, Evaluation at the Mean

	N = 500			N = 800		
	MAD	Med. Bias	RP 5%	MAD	Med. Bias	RP 5%
	A. High First Stage Signal/Noise, High Structural Signal/Noise					
Post-Double	0.330	-0.002	0.096	0.619	-0.176	0.068
Post-Single I	0.470	0.443	0.220	0.654	0.174	0.064
Post-Single II	1.402	1.402	0.676	5.436	5.436	1.000
Series I	0.489	0.442	0.200	0.971	0.936	0.204
Series II	0.418	-0.068	0.268	0.729	-0.141	0.120
Infeasible	0.234	0.051	0.104	0.255	-0.083	0.064
	B. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.312	-0.165	0.108	0.609	0.063	0.048
Post-Single I	0.311	0.015	0.088	0.616	0.239	0.052
Post-Single II	1.439	1.439	0.704	6.110	6.110	1.000
Series I	0.311	0.109	0.096	0.935	0.917	0.216
Series II	0.392	-0.200	0.236	0.685	0.024	0.104
Infeasible	0.247	-0.151	0.164	0.269	0.041	0.044
	C. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.504	-0.144	0.084	0.894	0.160	0.060
Post-Single I	0.524	0.293	0.124	1.028	0.582	0.052
Post-Single II	1.283	1.283	0.504	5.537	5.537	1.000
Series I	0.479	0.283	0.104	1.286	1.238	0.208
Series II	0.623	-0.267	0.240	0.912	0.063	0.112
Infeasible	0.276	-0.066	0.056	0.369	0.087	0.080
	D. Low First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.375	-0.108	0.080	0.767	0.113	0.036
Post-Single I	0.413	0.094	0.076	0.797	0.341	0.048
Post-Single II	1.483	1.483	0.620	6.191	6.191	1.000
Series I	0.401	0.165	0.068	1.107	1.018	0.120
Series II	0.515	-0.197	0.220	0.830	0.114	0.076
Infeasible	0.278	-0.089	0.076	0.361	0.087	0.036

Note: Results are based on 250 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for six different estimators of the function evaluated at the mean: the Post-Double proposed in this paper; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only; a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); an estimator that uses a small number of series terms (Series I); an estimator that uses a large number of series terms (Series II); and and infeasible estimator that is explicitly given the control function.

FIGURE 3. Simulation Table 3

Table 3. Simulation Results: High Dimensional Design, Average Derivative

	N = 500			N = 800		
	MAD	Med. Bias	RP 5%	MAD	Med. Bias	RP 5%
	A. High First Stage Signal/Noise, High Structural Signal/Noise					
Post-Double	0.056	0.027	0.124	0.057	-0.023	0.092
Post-Single I	0.300	-0.300	0.964	0.471	-0.471	1.000
Post-Single II	0.301	-0.301	0.968	0.473	-0.473	1.000
Series I	0.102	0.045	0.400	0.120	-0.016	0.420
Infeasible	0.059	0.045	0.140	0.057	0.003	0.048
	B. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.097	-0.095	0.356	0.058	-0.031	0.120
Post-Single I	0.365	-0.365	1.000	0.365	-0.365	0.984
Post-Single II	0.365	-0.365	1.000	0.365	-0.365	0.988
Series I	0.106	-0.070	0.464	0.099	-0.010	0.380
Infeasible	0.079	-0.079	0.328	0.053	-0.015	0.100
	C. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.081	-0.063	0.084	0.088	0.017	0.064
Post-Single I	0.372	-0.372	0.972	0.429	-0.429	0.980
Post-Single II	0.372	-0.372	0.972	0.432	-0.432	0.980
Series I	0.154	-0.053	0.408	0.179	0.050	0.428
Infeasible	0.069	-0.049	0.084	0.084	0.043	0.076
	D. Low First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.065	-0.033	0.084	0.072	0.035	0.076
Post-Single I	0.299	-0.299	0.940	0.315	-0.315	0.892
Post-Single II	0.299	-0.299	0.940	0.316	-0.316	0.896
Series I	0.130	-0.004	0.440	0.152	0.053	0.424
Infeasible	0.061	-0.019	0.100	0.069	0.046	0.120

Note: Results are based on 250 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for five different estimators of the average derivative: the Post-Double proposed in this paper; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only, a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); an estimator that includes every covariate (Series); and infeasible estimator that is explicitly given the control function.

FIGURE 4. Simulation Table 4

Table 4. Simulation Results: HighDimensional Design, Evaluation at the Mean

	N = 500			N = 800		
	MAD	Med. Bias	RP 5%	MAD	Med. Bias	RP 5%
	A. High First Stage Signal/Noise, High Structural Signal/Noise					
Post-Double	0.197	-0.013	0.104	0.238	0.031	0.060
Post-Single I	0.827	-0.827	0.780	0.343	0.230	0.096
Post-Single II	0.827	-0.827	0.780	0.343	0.241	0.096
Series I	0.395	0.002	0.372	0.527	-0.022	0.368
Infeasible	0.223	0.017	0.112	0.242	0.005	0.076
	B. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.214	-0.058	0.120	0.329	-0.142	0.052
Post-Single I	0.748	-0.748	0.684	0.293	0.030	0.044
Post-Single II	0.748	-0.748	0.684	0.293	0.030	0.044
Series I	0.388	0.032	0.404	0.721	-0.193	0.448
Infeasible	0.204	-0.016	0.120	0.334	-0.176	0.060
	C. High First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.302	-0.070	0.060	0.381	0.129	0.048
Post-Single I	0.885	-0.885	0.608	0.427	0.294	0.064
Post-Single II	0.885	-0.885	0.608	0.427	0.294	0.068
Series I	0.610	0.012	0.380	0.866	0.155	0.416
Infeasible	0.283	-0.048	0.084	0.375	0.110	0.040
	D. Low First Stage Signal/Noise, Low Structural Signal/Noise					
Post-Double	0.276	0.110	0.096	0.409	0.206	0.044
Post-Single I	0.565	-0.562	0.356	0.464	0.344	0.088
Post-Single II	0.569	-0.565	0.356	0.455	0.344	0.088
Series I	0.538	0.166	0.408	0.860	0.028	0.380
Infeasible	0.289	0.145	0.092	0.418	0.163	0.036

Note: Results are based on 250 simulation replications. The table reports median bias (Med. Bias), median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for five different estimators of the functional evaluated at the mean: the Post-Double proposed in this paper; a post-model selection estimator (Post-Single I) based on selecting terms with Lasso on the reduced form equation only, a post-model selection estimator (Post-Single II) based on selecting terms using Lasso on the outcome equation (Post-Single II); an estimator that includes every covariate (Series); and infeasible estimator that is explicitly given the control function.