cemmap
centre for microdata methods and practice

Program evaluation with high-dimensional data

Alexandre Belloni
Victor Chernozhukov
Iván Fernández-Val
Christian Hansen

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP57/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

PROGRAM EVALUATION WITH HIGH-DIMENSIONAL DATA

A. BELLONI, V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN

Abstract. We consider estimation of policy relevant treatment effects in a data-rich environment where there may be many more control variables available than there are observations. In addition to allowing many control variables, the setting we consider allows heterogeneous treatment effects, endogenous receipt of treatment, and function-valued outcomes. To make informative inference possible, we assume that reduced form predictive relationships are approximately sparse. That is, we require that the relationship between the covariates and the outcome, treatment status, and instrument status can be captured up to a small approximation error using a small number of controls whose identities are unknown to the researcher. This condition allows estimation and inference for a wide variety of treatment parameters to proceed after selection of an appropriate set of control variables formed by selecting controls separately for each reduced form relationship and then appropriately combining this set of reduced form predictive models and associated selected controls. We provide conditions under which post-selection inference is uniformly valid across a wide-range of models and show that a key condition underlying uniform validity of post-selection inference allowing for imperfect model selection is the use of approximately unbiased estimating equations. We illustrate the use of the proposed treatment effect estimation methods with an application to estimating the effect of 401(k) participation on accumulated assets.

Keywords: local average and quantile treatment effects, endogeneity, instruments, local effects of treatment on the treated, propensity score, LASSO

1. Introduction

The goal of many empirical analyses in economics is to understand the causal effect of some treatment such as participation in a government program on economic outcomes. Such analyses are often complicated by the fact that few economic treatments or government policies are randomly assigned. The lack of true random assignment has led to the adoption of a variety of quasi-experimental approaches to estimating treatment effects that are based on observational data. Such methods include instrumental variables (IV) methods in cases where treatment is not randomly assigned but there is some other external variable, such as eligibility for receipt of a government program or service, that is either randomly assigned or the researcher is willing to take as exogenous conditional on the right set of control variables. Another common approach

Date: First date: April 2013. This version: November 6, 2013. This is an abridged version of the paper intended for posting; a full version is available upon request. We gratefully acknowledge research support from the NSF. We are grateful to the seminar participants at University of Montreal, Summer NBER Institute, and the University of Illinois at Urbana-Champaign for helpful comments.

is to assume that the treatment variable itself may be taken as exogenous after conditioning on the right set of factors which leads to regression or matching based methods, among others, for estimating treatment effects.[1]

A practical problem empirical researchers must face when trying to estimate treatment effects is deciding what conditioning variables to include. When the treatment variable or instrument is not randomly assigned, a researcher must choose what needs to be conditioned on to make the argument that the instrument or treatment is exogenous plausible. Typically, economic intuition will suggest a set of variables that might be important to control for but will not identify exactly which variables are important or the functional form with which variables should enter the model. While less crucial to plausibly identifying treatment effects, the problem of selecting controls also arises in situations where the key treatment or instrumental variables are randomly assigned. In these cases, a researcher interested in obtaining precisely estimated policy effects will also typically consider including additional control variables to help absorb residual variation. As in the case where including controls is motivated by a desire to make identification of the treatment effect more plausible, one rarely knows exactly which variables will be most useful for accounting for residual variation. In either case, the lack of clear guidance about what variables to use presents the problem of selecting a set of controls from a potentially large set of variables including raw regressors available in the data as well as interactions and other transformations of these regressors.

In this paper, we consider estimation of the effect of an endogenous binary treatment, D, on an outcome, Y, in the presence of a binary instrumental variable, Z, in settings with very many controls, X. We allow for fully heterogeneous treatment effects and thus focus on estimation of causal quantities that are appropriate in heterogeneous effects settings such as the local average treatment effect (LATE) or the local quantile treatment effect (LQTE). We focus our discussion on the case where identification is obtained through the use of an instrumental variable but note that all results carry through to the case where the treatment is taken as exogenous after conditioning on sufficient controls simply by replacing the instrument with the treatment variable in the estimator and the formal results.

The methodology for estimating policy-relevant effects we consider allows for cases where the number of regressors, p, is much greater than the sample size, n. Of course, informative inference about causal parameters cannot proceed allowing for $p \gg n$ without further restrictions. We impose sufficient structure through the assumption that reduced form relationships such as $\mathrm{E}[D|X]$, $\mathrm{E}[Z|X]$, and $\mathrm{E}[Y|X]$ are approximately sparse. Intuitively, approximate sparsity imposes that these reduced form relationships can be represented up to a small approximation error as a linear combination, possibly inside of a known link function such as the logistic function, of a small number $s \ll n$ of the variables in X whose identities are *a priori* unknown to the researcher. This assumption allows us to use methods for estimating models in high-dimensional sparse settings

[1]There is a large literature about estimation of treatment effects. See, for example, the textbook treatments in Angrist and Pischke (2008) or Wooldridge (2010) and the references therein for discussion from an economics perspective.

that are known to have good prediction properties to estimate the fundamental reduced form relationships. We may then use these estimated reduced form quantities as inputs to estimating the causal parameters of interest. Approaching the problem of estimating treatment effects within this framework allows us to accommodate the realistic scenario in which a researcher is unsure about exactly which confounding variables or transformations of these confounds are important and so must search among a broad set of controls.

Valid inference following model selection is non-trivial as direct application of usual inference procedures following model selection does not provide valid inference about causal parameters even in low-dimensional settings, such as when there is only a single control, unless one assumes sufficient structure on the model that perfect model selection is possible. Such structure is very restrictive and seems unlikely to be satisfied in many economic applications. For example, a typical condition that allows perfect model selection in a linear model is to assume that all but a small number of coefficients are exactly zero and that the non-zero coefficients are all large enough that they can be distinguished from zero with probability very near one in finite samples. Such a condition rules out the possibility that there may be some variables which have moderate, but non-zero, partial effects. Ignoring such variables may lead to only a small loss in predictive performance while also producing a non-ignorable omitted variables bias that has a substantive impact on estimation and inference regarding individual model parameters. For further discussion, see Leeb and Pötscher (2008a; 2008b) and Pötscher (2009).

A key contribution of our paper is providing inferential procedures for key parameters used in program evaluation that are theoretically valid within approximately sparse models allowing for imperfect model selection. Our procedures build upon the insights in Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012) who demonstrate that valid inference about low-dimensional structural parameters can proceed following model selection, allowing for model selection mistakes, under two key conditions. First, estimation should be based upon "orthogonal" moment conditions that are first-order insensitive to changes in the values of the nuisance parameters. Specifically, if the target parameter value α_0 is idenitifed via the moment condition

$$\mathrm{E}_P \psi(W, \alpha_0, h_0) = 0, \tag{1}$$

where h_0 is to be estimated via a post-model-selection or regularization method, one needs to use a moment function, ψ, such that the moment condition is orthogonal with respect to perturbations of h around h_0. More formally, the moment conditions should satisfy

$$\partial_h [\mathrm{E}_P \psi(W, \alpha_0, h)]_{h=h_0} = 0 \tag{2}$$

where ∂h computes the functional derivative operator with respect to h. Second, one needs to use a model selection procedure that keeps model selection errors "moderately" small.

The orthogonality condition embodied in (2) has a long history in statistics and econometrics. For example, this type of orthogonality was used by Neyman (1979) in low-dimensional settings to deal with crudely estimated parametric nuisance parameters. To the best of our knowledge, Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012)

were the first to use this property in the $p \gg n$ setting where they used it to allow Lasso or post-Lasso estimation of the nuisance function h_0, the optimal instrument, in a linear instrumental variables model with many instruments. Using estimators based upon moment conditions with this low-bias property insures that crude estimation of h_0 via post-selection or other regularization methods has an asymptotically negligible effect on the estimation of α_0. Another paper in the $p \gg n$ setting where this approach is exploited is Belloni, Chernozhukov, and Hansen (2011) which proposes a double-selection method to construct an "orthogonal" moment equation whose use allows one to obtain valid inference on the parameters of the linear part of the partially linear model and on average treatment effects when treatment is exogenous conditional on observables.[2] In the general endogenous treatment effects setting we consider in this paper, such moment conditions can be found as efficient influence functions for certain reduced form parameters as in Hahn (1998). Moreover, our analysis allows for function-valued outcomes. As a result, the parameters of interest α_0 are themselves function-valued; i.e. they can carry an index. We illustrate how these efficient influence functions coupled with methods developed for forecasting in high-dimensional sparse models can be used to estimate and obtain valid inferential statements about a variety of structural/treatment effects. We formally demonstrate uniform in P validity of the resulting inference within a broad class of approximately sparse models including models where perfect model selection is theoretically impossible.

In establishing our main theoretical results, we consider variable selection for functional response data using ℓ_1-penalized methods. As examples, note that functional response data arises when one is interested in LQTE at not just a single quantile but across a range of quantile indices or when one is interested in how $1(Y \leq u)$ relates to treatment across a range of threshold values u. Considering such functional response data allows us to provide a unified inference procedure that allows for inference to be drawn about interesting quantities such as distributional effects of treatment as well as simpler objects such as the LQTE at a single quantile. Demonstrating that the developed methods provide uniformly valid inference for functional response data in a high-dimensional setting allowing for model selection mistakes is a main theoretical contribution of the present paper. Our result builds upon the work of Belloni and Chernozhukov (2011b) who provide rates of convergence for variable selection when one is interested in estimating the quantile regression process with exogenous variables. More generally, this theoretical work complements and extends the rapidly growing set of results for ℓ_1-penalized estimation methods; see, for example, Frank and Friedman (1993); Tibshirani (1996); Fan and Li (2001); Zou (2006); Candès and Tao (2007); Meinshausen and Yu (2009); Huang, Horowitz, and Ma (2008); Bickel, Ritov, and Tsybakov (2009); Huang, Horowitz, and Wei (2010); Belloni and Chernozhukov (2013); Bickel, Ritov, and Tsybakov (2009); Belloni, Chen, Chernozhukov, and Hansen (2012); van de Geer (2008); Bach (2010); Belloni, Chernozhukov, and Wei (2013); Belloni and Chernozhukov (2011a); Belloni, Chernozhukov, and Kato (2013); Kato (2011); and the references therein. We

[2]Farrell (2013) also builds upon Belloni, Chernozhukov, and Hansen (2011) focusing on the problem of estimating average treatment effects when treatment is exogenous conditional on observables.

also demonstrate that a simple weighted bootstrap procedure can be used to produce asymptotically valid inference statements which should aid in practical implementation of our proposed inference methods.

We illustrate the use of our methods by estimating the effect of 401(k) participation on measures of accumulated assets as in Chernozhukov and Hansen (2004).[3] Similar to Chernozhukov and Hansen (2004), we provide estimates of LATE and LQTE across a range of quantiles. We differ from this previous work by using the high-dimensional methods developed in this paper to allow ourselves to consider a much broader set of control variables than have previously been considered. We find that 401(k) participation has a small impact on accumulated financial assets at low quantiles while appearing to have a much larger impact at high quantiles. Interpreting the quantile index as "preference for savings" as in Chernozhukov and Hansen (2004), this pattern suggests that 401(k) participation has little causal impact on the accumulated financial assets of those with low desire to save but a much larger impact on those with stronger preferences for saving.[4] It is interesting that these results are quite similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of control variables.

1.1. **Notation.** A random variable W lives on the probability space $(S, \mathcal{S}, P)$. We have i.i.d. copies $(W_i)_{i=1}^n$ of W, the data. The data live on the probability space $(A, \mathcal{A}, \mathrm{P}_P)$, containing $\times_{i=1}^{\infty}(S, \mathcal{S}, P)$ as a subproduct. The probability space $(\Omega, \mathcal{A}, \mathrm{P}_P)$ will also carry i.i.d. copies of bootstrap multipliers $(\xi_i)_{i=1}^n$ which are independent of the data $(W_i)_{i=1}^n$. Note also that we use capital letters such as W to denote random elements and use the lower case letters such as w as fixed values that these random elements can take. We also use the standard empirical processes

$$\mathbb{G}_n(f) = \mathbb{G}_n(f(W)) = n^{-1/2} \sum_{i=1}^{n} \{f(W_i) - \mathrm{E}_P[f(W_i)]\}$$

indexed by a measurable class of functions $\mathcal{F} : S \mapsto \mathbb{R}$; see van der Vaart and Wellner (1996), Chapter 2.3. We denote by $\mathbb{P}_n$ the (random) empirical probability measure that assigns probability n^{-1} to each W_i. In what follows, we use $\|W\|_{\mathrm{P},q}$ to denote the $L^q(\mathrm{P})$ norm of a random variable W with law determined by P, and we use $\|W\|_{\mathbb{P}_n,q}$ to denote the empirical $L^q(\mathbb{P}_n)$ norm of a random variable with law determined by the empirical measure $\mathbb{P}_n$, i.e., $\|W\|_{\mathbb{P}_n,q} = (n^{-1} \sum_{i=1}^n \|W_i\|^q)^{1/q}$.

2. The Setting and The Target Parameters

2.1. **Observables and Reduced Form Parameters.** The observable random variables consist of Y_u, X, Z, and D, where Y_u is indexed by $u \in \mathcal{U}$. The observables are a random variable $W = ((Y_u)_{u\in\mathcal{U}}, X, Z, D)$. The variable $D \in \mathcal{D} = \{0, 1\}$ will indicate the receipt of a treatment

[3] See also Poterba, Venti, and Wise (1994; 1995; 1996; 2001); Benjamin (2003); and Abadie (2003) among others.

[4] Results in Chernozhukov and Hansen (2004) also suggest that there is little impact on total accumulated wealth at any quantile index suggesting that the results in the upper tail are largely due to substitutions from non-financial wealth to the tax advantaged 401(k) saving in financial assets.

or participation in a program. It will be typically endogenous; that is, we will typically view the treatment as assigned non-randomly. The instrumental variable $Z \in \mathcal{Z} = \{0, 1\}$ is a binary instrumental variable, such as an offer of participation, that is assumed to be exogenous conditional on observable covariates X. That is, we assume that we may treat the instrument as randomly assigned conditional on X. We denote the support of X by $\mathcal{X}$. The notions of exogeneity and endogeneity we employ are standard, but we state them below for clarity and completeness. We also restate standard conditions that are sufficient for a causal interpretation of our target parameters.

The random variable Y_u will be an outcome of interest. Allowing Y_u to be indexed is important for allowing treatment of functional data. For example, Y_u could represent an outcome falling short of a threshold, namely $Y_u = 1(Y \leqslant u)$, in the context of distributional analysis. In growth charts analysis, Y_u could be a height indexed by age u; and Y_u could be a health outcome indexed by a dosage u in dosage response studies. Our framework is tailored for such functional response data. The special case with no index is included by simply considering $\mathcal{U}$ as a singleton set.

We make use of two key types of reduced form parameters for estimating the structural parameters of interest – (local) treatment effects and related quantities. These reduced form parameters are defined as

$$\alpha_V(z) := \mathrm{E}_P[g_V(z, X)] \text{ for } z \in \{0, 1\} \text{ and } \gamma_V := \mathrm{E}_P[V], \tag{3}$$

where $z = 0$ or $z = 1$ are the fixed values of the random variable Z and the function g_V, mapping the support $\mathcal{Z}\mathcal{X}$ of vector (Z, X) to the real line $\mathbb{R}$, is defined as

$$g_V(z, x) := \mathrm{E}_P[V|Z = z, X = x]. \tag{4}$$

We use V to denote a target variable whose identity may change depending on the context such as $V = \mathbf{1}_d(D)Y_u$ or $V = \mathbf{1}_d(D)$ where $\mathbf{1}_d(D) := 1(D = d)$ is the indicator function.

The structural parameters we consider are smooth functionals of these reduced-form parameters. In our approach to estimating treatment effects, we estimate the key reduced form parameter $\alpha_V(z)$ using recent approaches to dealing with high-dimensional data coupled with using "low-bias" estimating equations. The low-bias property is crucial for dealing with the "non-regular" nature of penalized and post-selection estimators which do not admit linearizations except under very restrictive conditions. The use of regularization by model selection or penalization is in turn motivated by the desire to accommodate high-dimensional data.

2.2. **Target Structural Parameters – Local Treatment Effects.** The reduced form parameters defined in (3) are key because the structural parameters of interest are functionals of these elementary objects. The local average structural function (LASF) defined as

$$\theta_{Y_u}(d) = \frac{\alpha_{\mathbf{1}_d(D)Y_u}(1) - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\alpha_{\mathbf{1}_d(D)}(1) - \alpha_{\mathbf{1}_d(D)}(0)}, \quad d \in \{0, 1\} \tag{5}$$

underlies the formation of many commonly used treatment effects. The LASF identifies the average outcome for the group of *compliers*, individuals whose treatment state may be influenced by variation in the instrument, in the treated and non-treated states under standard assumptions;

see, e.g. Imbens and Angrist (1994). The local average treatment effect (LATE) is defined as the difference of the two values of the LASF:

$$\theta_{Y_u}(1) - \theta_{Y_u}(0). \tag{6}$$

The term local designates that this parameter does not measure the effect on the entire population but is local in the sense that it measures the effect for the subpopulation of compliers.

When there is no endogeneity, formally when $D \equiv Z$, the LASF and LATE become the average structural function (ASF) and average treatment effects (ATE). Thus, our results cover this situation as a special case. In this special case, the ASF and ATE are given by

$$\theta_{Y_u}(z) = \alpha_{Y_u}(z), \quad \theta_{Y_u}(1) - \theta_{Y_u}(0) = \alpha_{Y_u}(1) - \alpha_{Y_u}(0). \tag{7}$$

We also note that the impact of the instrument Z itself may be of interest since Z often encodes an offer of participation in a program. In this case, the parameters of interest are again simply the reduced form parameters $\alpha_{Y_u}(z)$ and $\alpha_{Y_u}(1)-\alpha_{Y_u}(0)$. Thus, the LASF and LATE are primary targets of interest in this paper with analysis of the ASF and ATE subsumed as special cases.

2.2.1. *Local Distribution and Quantile Treatment Effects.* Setting $Y_u = Y$ in (5) and (6) provides the conventional LASF and LATE. An important generalization arises by letting $Y_u = 1(Y \leqslant u)$ be the binary encoding of the outcome of interest falling below a threshold u. In this case, the family of effects

$$(\theta_{Y_u}(1) - \theta_{Y_u}(0))_{u\in\mathbb{R}}, \tag{8}$$

describe the local distributional treatment effects (LDTE). Similarly, we can look at the quantile transforms of the curves $u \mapsto \theta_{Y_u}(z)$,

$$\theta_Y^{\leftarrow}(\tau, z) := \inf\{u \in \mathbb{R} : \theta_{Y_u}(z) \geqslant \tau\}, \tag{9}$$

and examine the family of local quantile treatment effects (LQTE):

$$(\theta_Y^{\leftarrow}(\tau, 1) - \theta_Y^{\leftarrow}(\tau, 0))_{\tau\in(0,1)}. \tag{10}$$

2.3. **Target Structural Parameters – Local Treatment Effects on the Treated.** In addition to the local treatment effects given in Section 2.2, we may be interested in local treatment effects on the treated. The key object in defining local treatment effects on the treated is the local average structural function for the treated (LASF-T) which is defined by its two values:

$$\vartheta_{Y_u}(d) = \frac{\gamma_{\mathbf{1}_d(D)Y_u} - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\gamma_{\mathbf{1}_d(D)} - \alpha_{\mathbf{1}_d(D)(0)}}, \quad d \in \{0, 1\}. \tag{11}$$

These quantities identify the average outcome for the group of *treated compliers* in the treated and non-treated states under assumptions stated below. The local average treatment effect on the treated (LATE-T) introduced in Hong and Nekipelov (2010) is defined simply as the difference of two values of LASF-T:

$$\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0). \tag{12}$$

The LATE-T may be of interest because it measures the average treatment effect for *treated compliers*, namely the subgroup of compliers that actually receive the treatment.

When the treatment is assigned randomly given controls so we can take $D = Z$, the LASF-T and LATE-T become the average structural function for the treated (ASF-T) and average treatment effects on the treated (ATE-T). In this special case, the ASF-T and ATE-T are given by

$$\vartheta_{Y_u}(1) = \frac{\gamma_{\mathbf{1}_1(D)Y_u}}{\gamma_{\mathbf{1}_1(D)}}, \quad \vartheta_{Y_u}(0) = \frac{\gamma_{\mathbf{1}_0(D)Y_u} - \alpha_{Y_u}(0)}{\gamma_{\mathbf{1}_0(D)} - 1}, \quad \vartheta_{Y_u}(1) - \vartheta_{Y_u}(0); \tag{13}$$

and we can use our results to provide estimation and inference results for these quantities.

2.3.1. *Local Distribution and Quantile Treatment Effects on the Treated.* Local distributional treatment effects on the treated (LDTE-T) and local quantile treatment effects on the treated (LQTE-T) can also be defined. As in Section 2.2.1, we let $Y_u = 1(Y \leqslant u)$ be the binary encoding of an outcome of interest, Y, falling below a threshold u. The family of treatment effects

$$(\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0))_{u \in \mathbb{R}} \tag{14}$$

then describes the LDTE-T. We can also use the quantile transforms of the curves $u \mapsto \vartheta_{Y_u}(z)$,

$$\vartheta_Y^{\leftarrow}(\tau, z) := \inf\{u \in \mathbb{R} : \vartheta_{Y_u}(z) \geqslant \tau\}, \tag{15}$$

and define LQTE-T:

$$(\vartheta_Y^{\leftarrow}(\tau, 1) - \vartheta_Y^{\leftarrow}(\tau, 0))_{\tau \in (0,1)}. \tag{16}$$

Under conditional exogeneity LQTE and LQTE-T reduce the quantile treatment effects (QTE) (Koenker (2005)) and quantile treatment effects for the treated (QTE-T).

2.4. **Causal Interpretations for Structural Parameters.** The quantities discussed in Sections 2.2 and 2.3 have causal interpretations under standard conditions. To discuss these conditions, we use potential outcomes notation: Y_{u1} and Y_{u0} denote the potential outcomes under treatment states 1 and 0. These outcomes are not observed jointly, and we instead observe $Y_u = DY_{u1} + (1 - D)Y_{u0}$, where $D \in \mathcal{D} = \{0, 1\}$ is the random variable indicating participation or treatment state. Under exogeneity, D is assigned independently of the potential outcomes conditional on covariates X, i.e. $(Y_{u1}, Y_{u0}) \perp D \mid X$ a.s., where $\perp$ denotes statistical independence.

When exogeneity fails, D may depend on the potential outcomes. For example, people may drop out of a program if they think the program will not benefit them. In this case, instrumental variables are useful in creating quasi-experimental fluctuations in D that may identify useful effects. To provide identification in this setting, we assume the existence of an instrument Z, such as an offer of participation, that is assigned randomly conditional on observable covariates X. We further assume the instrument is binary. Let the random variables D_1 and D_0 indicate the participation decisions under the potential instrument states 1 and 0, respectively. These variables may in general depend on the potential outcomes. As with the potential outcomes, the participation decisions under both instrument states are not observed jointly. The realized participation decision is then given by $D = ZD_1 + (1 - Z)D_0$.

There are many causal quantities of interest for program evaluation. Chief among these are various structural averages

- average structural function (ASF): $\mathrm{E}_P[Y_{ud}]$,
- average structural function for the treated (ASF-T): $\mathrm{E}_P[Y_{ud} \mid D = 1]$,
- local average structural function (LASF): $\mathrm{E}_P[Y_{ud} \mid D_1 > D_0]$,
- local average structural function for the treated (LASF-T): $\mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1]$,

as well as effects derived from them such as

- average treatment effects (ATE): $\mathrm{E}_P[Y_{u1}] - \mathrm{E}_P[Y_{u0}]$,
- average treatment effect for the treated (ATE-T): $\mathrm{E}_P[Y_{u1} \mid D = 1] - \mathrm{E}_P[Y_{u0} \mid D = 1]$,
- local average treatment effect (LATE): $\mathrm{E}_P[Y_{u1} \mid D_1 > D_0] - \mathrm{E}_P[Y_{u0} \mid D_1 > D_0]$,
- local average treatment effect for the treated (LATE-T): $\mathrm{E}_P[Y_{u1} \mid D_1 > D_0, D = 1] - \mathrm{E}_P[Y_{u0} \mid D_1 > D_0, D = 1]$.

These causal quantities are the same as the stuctural parameters defined in Sections 2.2-2.3 under the following well-known sufficient condition.

Assumption 1 (Causal Interpretability)**.** The following conditions hold P-almost surely: (Exogeneity) $((Y_{u1}, Y_{u0})_{u \in \mathcal{U}}, D_1, D_0) \perp Z \mid X$; (First Stage) $\mathrm{E}_P[D_1 \mid X] \neq \mathrm{E}_P[D_0 \mid X]$; (Non-Degeneracy) $P(Z = 1 \mid X) \in (0, 1)$; (Monotonicity) $P(D_1 \geqslant D_0 \mid X) = 1$. ∎

This condition is much-used in the program evaluation literature. It also has an equivalent formulation in terms of a simultaneous equation model with a binary endogenous variable; see Vytlacil (2002) and Heckman and Vytlacil (1999). For a thorough discussion of this assumption, we refer to Imbens and Angrist (1994). Using this assumption, we present the following lemma which follows from results of Abadie (2003) and Hong and Nekipelov (2010) that both build upon the results of Imbens and Angrist (1994). The lemma shows that the parameters θ_{Y_u} and ϑ_{Y_u} defined earlier have a causal interpretation under Assumption 1. Therefore, our referring to them as structural/causal is justified under this condition.

Lemma 2.1 (Identification of Causal Effects)**.** *Under Assumption 1, for each $d \in \mathcal{D}$,*

$$\mathrm{E}_P[Y_{ud} \mid D_1 > D_0] = \theta_{Y_u}(d), \quad \mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1] = \vartheta_{Y_u}(d).$$

Furthermore, if D is exogenous, namely $D \equiv Z$ a.s., then

$$\mathrm{E}_P[Y_{ud} \mid D_1 > D_0] = \mathrm{E}_P[Y_{ud}], \quad \mathrm{E}_P[Y_{ud} \mid D_1 > D_0, D = 1] = \mathrm{E}_P[Y_{ud} \mid D = 1].$$

3. Estimation of Reduced-Form and Structural Parameters in a Data-Rich Environment

Recall that the key objects used in defining the structural parameters in Section 2 are the expectations

$$\alpha_V(z) = \mathrm{E}_P[g_V(z, X)] \text{ and } \gamma_V := \mathrm{E}[V], \tag{17}$$

where $g_V(z,X) = \mathrm{E}_P[V|Z=z,X]$ and V denotes a variable whose identity will change with the context. Specifically, we shall vary V over the set $\mathcal{V}_u$:

$$V \in \mathcal{V}_u := (V_{uj})_{j=1}^5 := \{Y_u, \mathbf{1}_d(D)Y_u, \mathbf{1}_d(D) : d \in \mathcal{D}\}. \tag{18}$$

Given the definition of $\alpha_V(z) = \mathrm{E}_P[g_V(z,X)]$, it is clear that $g_V(z,X)$ will play an important role in estimating $\alpha_V(z)$. A related function that will play an important role in forming a robust estimation strategy is the propensity score $m_Z : \mathcal{Z}\mathcal{X} \to \mathbb{R}$ defined by

$$m_Z(z,x) := \mathrm{P}_P[Z=z|X=x]. \tag{19}$$

We will denote other potential values for the functions g_V and m_Z by parameters g and m, respectively. A first approach to estimating $\alpha_V(z)$ is to try to recover g_V and m_Z directly using high-dimensional modelling and estimation methods.

As a second approach, we can further decompose g_V as

$$g_V(z,x) = \sum_{d=0}^{1} e_V(d,z,x) l_D(d,z,x), \tag{20}$$

where the regression functions e_V and l_D, mapping the support $\mathcal{D}\mathcal{Z}\mathcal{X}$ of (D,Z,X) to the real line, are defined by

$$e_V(d,z,x) := \mathrm{E}_P[V|D=d,Z=z,X=x] \quad \text{and} \tag{21}$$

$$l_D(d,z,x) := \mathrm{P}_P[D=d|Z=z,X=x]. \tag{22}$$

We shall denote other potential values for the functions e_V and l_D by parameters e and l. In this second approach, we can again use high-dimensional methods for modelling and estimating e_V and l_D, and we can then use relation (20) to obtain g_V. Given the resulting g_V and an estimate of m_Z obtained from using high-dimensional methods to model the propensity score, we can then recover $\alpha_V(z)$.

This second approach may be seen as a "special" case of the first. However, this approach could in fact be more principled than the first. For example, if we use linear or generalized linear[5] models to approximate each of the elements e_V, l_D and m_Z, then the implied approximations can strictly nest some coherent models such as the standard dummy endogenous variable model with normal disturbances. This strict nesting of coherent models is more awkward in the first approach which directly approximates g_V using linear or generalized linear forms. Indeed, the "natural" functional form for g_V is not of the linear or generalized linear form but rather is given by the affine aggregation of cross-products shown in (20). While these potential differences exist, we expect to see little quantitative difference between the estimates obtained via either approach if sufficiently flexible functional forms are used. For example, we see little difference between the two approaches in our empirical example.

[5]"Generalized linear" means "linear inside a known link function" in the context of the present paper.

3.1. **First Step: Modeling and Estimating Regression Function g_V, m_Z, l_D, and e_V in a Data-Rich Environment.** In this section, we elaborate the two strategies that we introduced above.

Strategy 1. We first discuss direct estimation of g_V and m_Z, which corresponds to the first strategy suggested in the previous subsection. Since the functions are unknown and potentially complicated, we use generalized linear combinations of a large number of control terms

$$f(X) = (f_j(X))_{j=1}^p, \tag{23}$$

to approximate g_V and m_Z. Specifically, we use

$$g_V(z,x) =: \Lambda_V[f(x,z)'\beta_V] + r_V(z,x), \tag{24}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \quad \beta_V := (\beta_V(0)', \beta_V(1)')', \tag{25}$$

and

$$m_Z(1,x) =: \Lambda_Z[f(x)'\beta_Z] + r_Z(x), \quad m_Z(0,x) = 1 - \Lambda_Z[f(x)'\beta_Z] - r_Z(x). \tag{26}$$

In these equations, $r_V(z,x)$ and $r_Z(x)$ are approximation errors, and the functions $\Lambda_V(f(x)'\beta_V(z))$ and $\Lambda_Z(f(x)'\beta_Z)$ are generalized linear approximations to the target functions $g_V(z,x)$ and $m_Z(1,x)$. The functions Λ_V and Λ_Z are taken to be known link functions Λ. The most common example is the linear link $\Lambda(u) = u$. When the response variables V, Z, and D are binary, we may also use the logistic link $\Lambda(u) = \Lambda_0(u) = e^u/(1+e^u)$ and its complement $1 - \Lambda_0(u)$ or the probit link $\Lambda(u) = \Phi(u) = (2\pi)^{-1}\int_{-\infty}^u e^{-z^2/2}dz$ and its complement $1 - \Phi(u)$. For clarity, we use links from the finite set $\mathcal{L} = \{\mathrm{Id}, \Phi, 1-\Phi, \Lambda_0, 1-\Lambda_0\}$ where Id is the identity (linear) link.

In order to allow for a flexible specification and incorporation of pertinent confounding factors, we allow for the dictionary of controls, denoted $f(X)$, to be "rich" in the sense that its dimension $p = p_n$ may be large relative to the sample size. Specifically, our results require only that

$$\log p = o(n^{1/3})$$

along with other technical conditions. High-dimensional regressors $f(X)$ could arise for different reasons. For instance, the list of available controls could be large, i.e. $f(X) = X$ as in e.g. Koenker (1988). It could also be that many technical controls are present; i.e. the list $f(X) = (f_j(X))_{j=1}^p$ could be composed of a large number of transformations of elementary regressors X such as B-splines, dummies, polynomials, and various interactions as, e.g., in Newey (1997), Tsybakov (2009), and Wasserman (2006). The functions f_j forming the dictionary can depend on n, but we suppress this dependence.

Having very many controls $f(X)$ creates a challenge for estimation and inference. A useful condition that makes it possible to perform constructive estimation and inference in such cases is termed approximate sparsity or simply sparsity. Sparsity imposes that there exist approximations of the form given in (24)-(26) that require only a small number of non-zero coefficients to render the approximation errors small relative to estimation error. More formally, sparsity relies on two

conditions. First, there must exist β_V and β_Z such that

$$\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{27}$$

That is, there are at most $s = s_n \ll n$ elements in the union of the support of β_Z with the union of the supports of β_V. Second, the sparsity condition requires that the size of the resulting approximation errors is small compared to the conjectured size of the estimation error; namely, for all $V \in \mathcal{V}$,

$$\{\mathrm{E}_P[r_V^2(Z,X)]\}^{1/2} + \{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{28}$$

Note that the size of the approximating model $s = s_n$ can grow with n just as in standard series estimation, subject to the rate condition

$$s^2 \log^3(p \vee n)/n \to 0.$$

This condition ensures that functions g_V and m_Z are estimable at the $o(n^{-1/4})$ rates and are used to derive asymptotic normality results for the structural and reduced-form parameter estimates. This condition can be substantially relaxed if sample splitting methods are used.

The high-dimensional-sparse-model framework outlined above extends the standard framework in the program evaluation literature which assumes both that the identities of the relevant controls are known and that the number of such controls s is much smaller than the sample size. Instead, we assume that there are many, p, potential controls of which at most s controls suffice to achieve a desirable approximation to the unknown functions g_V and m_Z; and we allow the identity of these controls to be unknown. Relying on this assumed sparsity, we use selection methods to choose approximately the right set of controls.

Current estimation methods that exploit approximate sparsity employ different types of regularization aimed at producing estimators that theoretically perform well in high-dimensional settings while remaining computationally tractable. Many widely used method are based on ℓ_1-penalization. The Lasso method is one such commonly used approach that adds a penalty for the weighted sum of the absolute values of model parameters to the usual objective function of an M-estimator. A related approach is the Post-Lasso method which performs re-estimation of the model after selection of variables by Lasso. These methods are discussed at length in recent papers and review articles; see, for example, Belloni, Chernozhukov, and Hansen (2013). Rather than provide specifics of these methods here, we specify detailed implementation algorithms in the Appendix.

In the following, we outline the general features of the Lasso method focusing on estimation of g_V. Given data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (V_i, f(Z_i, X_i))_{i=1}^n$, the Lasso estimator $\widehat{\beta}_V$ solves

$$\widehat{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^p} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] + \frac{\lambda}{n} \|\widehat{\Psi}\beta\|_1 \Big), \tag{29}$$

where $\widehat{\Psi} = \mathrm{diag}(\widehat{l}_1, \ldots, \widehat{l}_p)$ is a diagonal matrix of data-dependent penalty loadings, $M(y,t) = .5(y-t)^2$ in the case of linear regression, and $M(y,t) = 1(y=1)\log\Lambda(t) + 1(y=0)\log(1-\Lambda(t))$ in the case of binary regression. In the binary case, the link function Λ could be logistic or probit. The penalty level, λ, and loadings, $\widehat{l}_j$ $j = 1, ..., p$, are selected to guarantee good

theoretical properties of the method. We provide theoretical choices and further detail regarding implementation in Section 5. A key consideration in this paper is that the penalty level needs to be set to account for the fact that we will be simultaneously estimating potentially a *continuum* of Lasso regressions since our V varies over the list $\mathcal{V}_u$ with u varying over the index set $\mathcal{U}$.

The post-Lasso method uses $\widehat{\beta}_V$ solely as a model selection device. Specifically, it makes use of the labels of the regressors with non-zero estimated coefficients,

$$\widehat{I}_V = \text{support}(\widehat{\beta}_V).$$

The Post-Lasso estimator is then a solution to

$$\tilde{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^p} \Big(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] : \beta_j = 0, j \notin \widehat{I}_V \Big). \tag{30}$$

A main contribution of this paper is establishing that estimators $\widehat{g}_V(Z,X) = \Lambda(f(Z,X)'\bar{\beta}_V)$ of the regression function $g_V(Z,X)$, where $\bar{\beta}_V = \widehat{\beta}_V$ or $\bar{\beta}_V = \tilde{\beta}_V$, achieve the near oracle rate of convergence $\sqrt{(s \log p)/n}$ and maintain desirable theoretic properties allowing for a *continuum* of response variables.

Estimation of m_Z proceeds similarly. The Lasso estimator $\widehat{\beta}_Z$ and Post-Lasso estimators $\tilde{\beta}_Z$ are defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (Z_i, f(X_i))_{i=1}^n$. As with the estimator $\widehat{g}_V(Z,X)$, the estimator $\widehat{m}_Z(1,X) = \Lambda_Z(f(X)'\bar{\beta}_Z)$ of $m_Z(X)$, with $\bar{\beta}_Z = \widehat{\beta}_Z$ or $\bar{\beta}_Z = \tilde{\beta}_Z$, achieves the near oracle rate or convergence $\sqrt{(s \log p)/n}$ and has other good theoretic properties. The estimator of $\widehat{m}_Z(0,X)$ is then given by $1 - \widehat{m}_Z(1,X)$.

Strategy 2. The second strategy we consider involves modeling and estimating m_Z as above via (26) while modeling g_V via its disaggregation into parts e_V and l_D via (20). We model each of the unknown parts[6] of e_V and l_D using the same approach as in Strategy 1. Specifically, we model the conditional expectation of V given D, Z, and X by

$$e_V(d,z,x) =: \Gamma_V[f(d,z,x)'\theta_V] + \varrho_V(d,z,x), \tag{31}$$

$$f(d,z,x) := ((1-d)f(z,x)', df(z,x)')', \ \theta_V := (\theta_V(0)', \theta_V(1)')'. \tag{32}$$

We model the conditional probability of D taking on 1 or 0, given Z and X by

$$l_D(1,z,x) =: \Gamma_D[f(z,x)'\theta_D] + \varrho_D(z,x), \tag{33}$$

$$l_D(0,z,x) = 1 - \Gamma_D[f(z,x)'\theta_D] - \varrho_D(z,x), \tag{34}$$

$$f(z,x) := ((1-z)f(x)', zf(x)')', \tag{35}$$

$$\theta_D := (\theta_V(0,0)', \theta_V(0,1)', \theta_V(1,0)', \theta_V(1,1)')'. \tag{36}$$

Here $\varrho_V(d,z,x)$ and $\varrho_D(z,x)$ are approximation errors, and the functions $\Gamma_V(f(X)'\theta_V(d,z,x))$ and $\Gamma_D(f(X)'\theta_V(z,x))$ are generalized linear approximations to the target functions $e_V(d,z,x)$ and $l_D(1,z,x)$. The functions Γ_V and Γ_D are taken to be known link functions $\Lambda \in \mathcal{L}$ as in the previous strategy.

[6]Upon conditioning on $D = d$ some parts become known; e.g., $e_{1_d(D)Y}(d',x,z) = 0$ if $d \neq d'$ and $e_{1_d(D)}(d',x,z) = 1$ if $d = d'$.

As in the first strategy, we maintain approximate sparsity in the modelling framework. We assume that there exist β_Z, θ_V and θ_D such that

$$\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \leqslant s. \tag{37}$$

That is, we assume that there are at most $s = s_n \ll n$ elements in the union of the support of β_Z, the support of θ_D, and the union of the supports of θ_V where the union is over variables $V \in \mathcal{V} = (\mathcal{V}_u, u \in \mathcal{U})$. The sparsity condition also requires the size of the approximation errors to be small compared to the conjectured size of the estimation error: For all $V \in \mathcal{V} = (\mathcal{V}_u, u \in \mathcal{U})$, we assume

$$\{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} + \{\mathrm{E}_P[\varrho_V^2(D,Z,X)]\}^{1/2} + \{\mathrm{E}_P[\varrho_D^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{38}$$

Note that the size of the approximating model $s = s_n$ can grow with n just as in standard series estimation as long as $s^2 \log^3(p \vee n)/n \to 0$.

We proceed with estimation of e_V and l_D analogously to the approach outlined in Strategy 1. The Lasso estimator $\widehat{\theta}_V$ and Post-Lasso estimators $\tilde{\theta}_V$ are defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (V_i, f(D_i, Z_i, X_i))_{i=1}^n$ and the link function $\Lambda = \Lambda_V$. The estimators $\widehat{e}_V(D,Z,X) = \Lambda(f(D,Z,X)'\bar{\theta}_V)$, with $\bar{\theta}_V = \widehat{\theta}_V$ or $\bar{\theta}_V = \widehat{\theta}_V$, have near oracle rates or convergence, $\sqrt{(s \log p)/n}$, and other desirable properties. The Lasso estimator $\widehat{\theta}_D$ and Post-Lasso estimators $\tilde{\theta}_D$ are also defined analogously to $\widehat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (D_i, f(Z_i, X_i))_{i=1}^n$ and the link function $\Lambda = \Lambda_D$. Again, the estimators $\widehat{l}_D(Z,X) = \Lambda(f(Z,X)'\bar{\theta}_D)$ of $l_D(Z,X)$, where $\bar{\theta}_D = \widehat{\theta}_D$ or $\bar{\theta}_D = \tilde{\theta}_D$, have good theoretical properties including the near oracle rate of convergence, $\sqrt{(s \log p)/n}$. The resulting estimator for $g_V(z,X)$ is then given by

$$\widehat{g}_V(z,x) = \sum_{d=0}^{1} \widehat{e}_V(d,z,x)\widehat{l}_D(d,z,x). \tag{39}$$

3.2. **Second Step: Robust Estimation of Reduced-Form Parameters $\alpha_V(z)$.** Estimation of the key quantities $\alpha_V(z)$ will make heavy use of "low-bias" moment functions as defined in (2). These moment functions are closely tied to efficient influence functions, where efficiency is in the sense of locally minimax semi-parametric efficiency. The use of these functions will deliver robustness with respect to the irregularity of the post-selection and penalized estimators needed to manage high-dimensional data. The use of these functions also automatically delivers semi-parametric efficiency for estimating and performing inference on the reduced-form parameters and their smooth transformations – the structural parameters.

The efficient influence function and low-bias moment function for $\alpha_V(z)$ for $z \in \mathcal{Z} = \{0,1\}$ are given respectively by

$$\psi_{V,z}^{\alpha}(W) := \psi_{V,z,g_V,m_Z}^{\alpha}(W, \alpha_V(z)) \quad \text{and} \tag{40}$$

$$\psi_{V,z,g,m}^{\alpha}(W,\alpha) := \frac{1(Z=z)(V - g(z,X))}{m(z,X)} + g(z,X) - \alpha. \tag{41}$$

The efficient influence function was derived by Hahn (1998); they were also used by (Cattaneo, 2010) in the series context (with $p \ll n$) and (Rothe and Firpo, 2013) in the kernel contexts.

The efficient influence function and the moment function for γ_V are trivially given by

$$\psi_V^\gamma(W) := \psi_V^\gamma(W, \gamma_V), \text{ and } \psi_V^\gamma(W, \gamma) := V - \gamma. \tag{42}$$

We then define the estimator of the reduced-form parameters $\alpha_V(z)$ and $\gamma_V(z)$ as solutions $\alpha = \widehat{\alpha}_V(z)$ and $\gamma = \widehat{\gamma}_V$ to the equations

$$\mathbb{E}_n[\psi^\alpha_{V,z,\widehat{g}_V,\widehat{m}_Z}(W, \alpha)] = 0, \quad \mathbb{E}_n[\psi_V^\gamma(W, \gamma)] = 0, \tag{43}$$

where $\widehat{g}_V(z, x)$ and $\widehat{m}_Z(z, x)$ are constructed as in the previous section. Note that $\widehat{g}_V(z, x)$ may be constructed via either Strategy 1 or Strategy 2. We apply this procedure to each variable name $V \in \mathcal{V}_u$ and obtain the estimator[7]

$$\widehat{\rho}_u := \big(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\big)_{V \in \mathcal{V}_u} \quad \text{of} \quad \rho_u := \big(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\big)_{V \in \mathcal{V}_u}. \tag{44}$$

The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with dimension $d_\rho = 15$.

In the next section, we formally establish a principal result which shows that

$$\sqrt{n}(\widehat{\rho}_u - \rho_u) \rightsquigarrow N(0, \mathrm{Var}_P(\psi_u^\rho)), \quad \psi_u^\rho := (\{\psi^\alpha_{V,0}, \psi^\alpha_{V,1}, \psi_V^\gamma\})_{V \in \mathcal{V}_u}, \tag{45}$$
$$\text{uniformly in } P \in \mathcal{P}_n,$$

where $\mathcal{P}_n$ is a rich set of data generating processes P. The notation "$\rightsquigarrow$ uniformly in $P \in \mathcal{P}_n$" is defined formally in the Appendix and can be read as "approximately distributed as uniformly in $P \in \mathcal{P}_n$." This usage corresponds to the usual notion of asymptotic distribution extended to handle uniformity in P. Here $\mathcal{P}_n$ is a "rich" set of data generating processes P which includes cases where perfect model selection is impossible theoretically.

We then denote all the reduced form estimators and the estimands as

$$\widehat{\rho} = (\widehat{\rho}_u)_{u \in \mathcal{U}} \quad \text{and} \quad \rho = (\rho_u)_{u \in \mathcal{U}},$$

giving rise to the empirical reduced-form process $\widehat{\rho}$ and a reduced form functional ρ. We establish that $\sqrt{n}(\widehat{\rho} - \rho)$ is asymptotically Gaussian: In $\ell^\infty(\mathcal{U})^{d_\rho}$, we have

$$\sqrt{n}(\widehat{\rho} - \rho) \rightsquigarrow Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}, \text{ uniformly in } P \in \mathcal{P}_n \tag{46}$$

where $\mathbb{G}_P$ denotes the P-Brownian bridge (van der Vaart and Wellner, 1996). This result contains (45) as a special case and again allows $\mathcal{P}_n$ to be a "rich" set of data generating processes P that includes cases where perfect model selection is impossible theoretically. Importantly, this result verifes the functional central limit theorem applies for the reduced-form estimators in the presence of possible model selection mistakes.

Since some of our objects of interest are complicated, inference can be facilitated by the multiplier bootstrap. We define a bootstrap draw of $\widehat{\rho}^* = (\widehat{\rho}^*_u)_{u \in \mathcal{U}}$ via

$$\sqrt{n}(\widehat{\rho}^*_u - \widehat{\rho}_u) = n^{-1/2} \sum_{i=1}^n \xi_i \widehat{\psi}^\rho_u(W_i). \tag{47}$$

[7]By default notation, $(a_j)_{j \in \mathcal{J}}$ returns a column vector produced by stacking components together in some consistent order.

Here $(\xi_i)_{i=1}^n$ are i.i.d. copies of ξ which are independently distributed from the data $(W_i)_{i=1}^n$ and whose distribution does not depend on P. We also impose that

$$\mathrm{E}[\xi] = 0, \quad \mathrm{E}[\xi^2] = 1, \quad \mathrm{E}[\exp(|\xi|)] < \infty.$$

Examples of ξ include (a) $\xi = \mathcal{E} - 1$, where $\mathcal{E}$ is standard exponential random variable, (b) $\xi = \mathcal{N}$, where $\mathcal{N}$ is standard normal random variable, and (c) $\xi = \mathcal{N}/\sqrt{2} + (\mathcal{N}^2 - 1)/2$, where $\mathcal{N}$ is standard normal random variable.[8] Method (a), (b), and (c) correspond respectively to a Bayesian bootstrap (e.g., Chamberlain and Imbens (2003)), a Gaussian multiplier method (e.g, van der Vaart and Wellner (1996)), and a wild bootstrap method (Mammen, 1993).[9] $\widehat{\psi}_u^\rho$ in (47) are estimates of the influence function ψ_u^ρ defined via the plug-in rule:

$$\widehat{\psi}_u^\rho = (\widehat{\psi}_V^\rho)_{V \in \mathcal{V}_u}, \quad \widehat{\psi}_V^\rho(W) := \{\psi^\alpha_{V,0,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(0)), \psi^\alpha_{V,1,\widehat{g}_V,\widehat{m}_Z}(W, \widehat{\alpha}_V(1)), \psi^\gamma_V(W, \widehat{\gamma}_V)\}. \quad (48)$$

Note that this bootstrapping is computationally efficient since it does not involve recomputing the influence functions $\widehat{\psi}_u$. Each new draw of $(\xi_i)_{i=1}^n$ generates a new draw of $\widehat{\rho}^*$ holding the data and the estimates of the influence functions fixed. Note that this method simply amounts to resampling the first-order approximations to the estimators.

We establish that that the bootstrap law $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}_u)$ is uniformly asymptotically valid: In the metric space $\ell^\infty(\mathcal{U})^{d_\rho}$, both unconditionally and conditionally on the data,

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) \rightsquigarrow_B Z_P, \text{ uniformly in } P \in \mathcal{P}_n,$$

where $\rightsquigarrow_B$ denotes the convergence of the bootstrap law conditional on the data, as defined in the Appendix.

3.3. **Step 3: Robust Estimates of the Structural Parameters.** All structural parameters we consider take the form of smooth transformations of reduced form parameters:

$$\Delta = (\Delta_q)_{q \in \mathcal{Q}}, \text{ where } \Delta_q := \phi(\rho)(q), \ q \in \mathcal{Q}. \quad (49)$$

The structural parameters may themselves carry an index $q \in \mathcal{Q}$; for example, the structural quantile treatment effects are indexed by the quantile index. This formulation includes as special cases all the structural functions we previously mentioned. We estimate these quantities by the plug-in rule. We establish the good asymptotic behavior of these estimators and the validity of the bootstrap as a corollary from the results outlined in Section 3.2 and the functional delta method.

For application of the functional delta method, we require that the functionals be Hadamard differentiable – tangential to a subset that contains realizations of Z_P for all $P \in \mathcal{P}_n$ – with

[8] We do not consider the nonparametric bootstrap, which corresponds to using multinomial weights, to reduce the length of the paper; but we note that it is possible to show that it is also valid in the present setting.

[9] The motivation for method (c) is that it is able to match 3 moments since $\mathrm{E}[\xi^2] = \mathrm{E}[\xi^3] = 1$. Methods (a) and (b) do not satisfy this property since $\mathrm{E}[\xi^2] = 1$ but $\mathrm{E}[\xi^3] \neq 1$ for these approaches.

derivative map $h \mapsto \phi'_\rho(h) = (\phi'_{\rho,q}(h))_{q\in\mathcal{Q}}$. We define the estimators and their bootstrap versions as

$$\widehat{\Delta}_q := \phi\left(\widehat{\rho}\right)(q), \quad \widehat{\Delta}^*_q := \phi\left(\rho^*_u\right)(q). \tag{50}$$

We establish that these estimators are asymptotically Gaussian

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n, \tag{51}$$

and that the bootstrap consistently estimates their large sample distribution uniformly in $P \in \mathcal{P}_n$:

$$\sqrt{n}(\widehat{\Delta}^* - \Delta^*) \rightsquigarrow_B \phi'_\rho(Z_P), \text{ uniformly in } P \in \mathcal{P}_n. \tag{52}$$

These results can be used as an ingredient to standard construction of simultaneous confidence bands on Δ.

4. Theory of Estimation and Inference on Local Treatment Effects Functionals

Consider fixed sequences of positive numbers $\delta_n \searrow 0$, $\epsilon_n \searrow 0$, $\Delta_n \searrow 0$, $\ell_n \to \infty$, and $1 \leqslant K_n < \infty$ and positive constants c, C, and $c' < 1/2$ which will not vary with P. P is allowed to vary in the set $\mathcal{P}_n$ of probability measures, termed "data-generating processes", where $\mathcal{P}_n$ is typically a weakly increasing in n set.

Assumption 2. (Basic Assumptions) (i) For each $n \geqslant 1$, our data will consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X, Z, D)$ defined on the probability space $(S, \mathcal{S}, P)$, where $P \in \mathcal{P}_n$, and the collection $(Y_u)_{u\in\mathcal{U}}$ is suitably measurable, namely image-admissible Suslin. Let

$$V_u := (V_{uj})_{j\in\mathcal{J}} := \{Y_u, \mathbf{1}_d(D)Y_u, \mathbf{1}_d(D) : d \in \mathcal{D}\}$$

where $\mathcal{J} = \{1, ..., 5\}$ and $\mathcal{V} = (V_u)_{u\in\mathcal{U}}$. (ii) For $\mathcal{P} := \cup_n \mathcal{P}_n$, the map $u \mapsto Y_u$ obeys the uniform continuity property:

$$\lim_{\epsilon\searrow 0} \sup_{P\in\mathcal{P}} \sup_{d_{\mathcal{U}}(u,\bar{u})\leqslant\epsilon} \|Y_u - Y_{\bar{u}}\|_{P,2} = 0, \ \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{u\in\mathcal{U}} |Y_u|^{2+c} \leqslant \infty,$$

for each $j \in \mathcal{J}$, where the supremum is taken over $u, \bar{u} \in \mathcal{U}$, and $\mathcal{U}$ is a totally bounded metric space equipped with the metric $d_{\mathcal{U}}$. The uniform ϵ covering entropy of $(Y_u, u \in \mathcal{U})$ is bounded by $C \log(\mathrm{e}/\epsilon) \vee 0$. (iii) For each $P \in \mathcal{P}$, the conditional probability of $Z = 1$ given X is bounded away from zero or one: $P(c' \leqslant m_Z(z, X) \leqslant 1 - c') = 1$; the instrument Z has a non-trivial impact on D, namely $P(c' \leqslant l_D(1, 1, X) - l_D(1, 0, X)) = 1$; and the regression function g_V is bounded $\|g_V\|_{P,\infty} < \infty$ for all $V \in \mathcal{V}$. ∎

This assumption implies that the set of functions $(\psi_u)_{u\in\mathcal{U}}$, where $\psi^\rho_u := (\{\psi^\alpha_{V,0}, \psi^\alpha_{V,1}, \psi^\gamma_V\})_{V\in\mathcal{V}_u}$, is P-Donsker uniformly in $\mathcal{P}$. That is, it implies

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}, \tag{53}$$

where

$$Z_{n,P} := (\mathbb{G}_n\psi^\rho_u)_{u\in\mathcal{U}} \text{ and } Z_P := (\mathbb{G}_P\psi^\rho_u)_{u\in\mathcal{U}}, \tag{54}$$

with $\mathbb{G}_P$ denoting the P-Brownian bridge (van der Vaart and Wellner, 1996), and Z_P having bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{u\in\mathcal{U}} \|Z_P(u)\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u,\tilde{u})\leqslant\varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| = 0. \tag{55}$$

Other assumptions will be specific to the strategy taken.

Assumption 3 (Approximate Sparsity for Strategy 1)**.** Under each $P \in \mathcal{P}_n$ and for each $n \geqslant n_0$, uniformly for all $V \in \mathcal{V}$ the following hold: (i) approximations (24)-(26) hold with the link functions Λ_V and Λ_Z belonging to the set $\mathcal{L}$, the sparsity condition holding, $\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s$, approximation errors satisfying $\|r_V\|_{P,2} + \|r_Z\|_{P,2} \leqslant \delta_n n^{-1/4}$, $\|r_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \leqslant \epsilon_n$, and the sparsity index and the number of terms p in vector $f(X)$ obeying $s^2 \log^3(p \vee n)/n \leqslant \delta_n$. (ii) There are estimators $\bar{\beta}_V$ and $\bar{\beta}_Z$ such that, with probability no less than $1 - \Delta_n$, estimation errors satisfy $\|f(Z,X)'(\bar{\beta}_V - \beta_V)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$, $K_n\|\bar{\beta}_Z - \beta_Z\|_1 + K_n\|\bar{\beta}_Z - \beta_Z\|_1 \leqslant \delta_n$; the estimators are sparse such that $\|\bar{\beta}_V\|_0 + \|\bar{\beta}_Z\|_0 \leqslant Cs$; and the empirical and populations norms induced by the Gram matrix formed by $(f(X_i))_{i=1}^n$ are equivalent on sparse subsets, $\sup_{\|\delta\|_0\leqslant\ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2}/\|f(X)'\delta\|_{P,2} - 1| \leqslant \delta_n$. (iii) The following boundedness conditions hold: $\|\|f(X)\|_\infty\|_{P,\infty} \leqslant K_n$ and $\|V\|_{P,\infty} \leqslant C$. ∎

Comment 4.1. These conditions are simple intermediate-level conditions which encode both the approximate sparsity of the models as well as some reasonable behavior on the sparse estimators of m_Z and g_V. Sufficient conditions for the equivalence between empirical and population norms are given in Belloni, Chernozhukov, and Hansen (2011). The boundedness conditions are made to simplify arguments, and they could be removed at the cost of more complicated proofs and more stringent side conditions. ∎

Assumption 4 (Approximate Sparsity for Strategy 2)**.** Under each $P \in \mathcal{P}_n$ and for all $n \geqslant n_0$, uniformly for all $V \in \mathcal{V}$ the following hold: (i) Approximations (31)-(33) and (26) apply with the link functions Γ_V, Γ_D and Λ_Z belonging to the set $\mathcal{L}$, the sparsity condition $\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \leqslant s$ holding, approximation errors satisfying $\|\varrho_D\|_{P,2} + \|\varrho_V\|_{P,2} + \|r_Z\|_{P,2} \leqslant \delta_n n^{-1/4}$ and $\|\varrho_D\|_{P,\infty} + \|\varrho_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \leqslant \epsilon_n$, and the sparsity index s and the number of terms p obeying $s^2 \log^3(p \vee n)/n \leqslant \delta_n$. (ii) There are estimators $\bar{\theta}_V$, $\bar{\theta}_D$, and $\bar{\beta}_Z$ such that, with probability no less than $1 - \Delta_n$, estimation errors satisfy $\|f(D,Z,X)'(\bar{\theta}_V - \theta_V)\|_{\mathbb{P}_n,2} + \|f(Z,X)'(\bar{\theta}_D - \theta_D)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$ and $K_n\|\bar{\theta}_V - \theta_V\|_1 + K_n\|\bar{\theta}_D - \theta_D\|_1 + K_n\|\bar{\beta}_Z - \beta_Z\|_1 \leqslant \epsilon_n$; the estimators are sparse such that $\|\bar{\theta}_V\|_0 + \|\bar{\theta}_D\|_0 + \|\bar{\beta}_Z\|_0 \leqslant Cs$; and the empirical and populations norms induced by the Gram matrix formed by $(f(X_i))_{i=1}^n$ are equivalent on sparse subsets, $\sup_{\|\delta\|_0\leqslant\ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2}/\|f(X)'\delta\|_{P,2} - 1| \leqslant \delta_n$. (iii) The following boundedness conditions hold: $\|\|f(X)\|_\infty\|_{P,\infty} \leqslant K_n$ and $\|V\|_{P,\infty} \leqslant C$. ∎

Under the stated assumptions, the empirical reduced form process $\widehat{Z}_{n,P} := \sqrt{n}(\widehat{\rho} - \rho)$ defined by (44) obeys the following laws.

Theorem 4.1 (**Uniform Gaussianity of the Reduced-Form Parameter Process**). *Under Assumptions 2 and 3 or 2 and 4 holding, the reduced-form empirical process admits a linearization, namely*

$$\widehat{Z}_{n,P} := \sqrt{n}(\widehat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{56}$$

The process is also asymptotically Gaussian, namely

$$\widehat{Z}_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{57}$$

where Z_P is defined in (54) and its paths obey the property (55) .

Another main result of this section shows that the bootstrap law

$$\widehat{Z}^*_{n,P} = \sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$$

provides a valid approximation to the large sample law of $\sqrt{n}(\widehat{\rho} - \rho)$.

Theorem 4.2 (**Validity of Weighted Bootsrap for Inference on Reduced-Form Parameters**). *Under Assumptions 2 and 3 or 2 and 4, the weighted bootstrap consistently approximates the large sample laws Z_P of $Z_{n,p}$ uniformly in $P \in \mathcal{P}_n$, namely,*

$$\widehat{Z}^*_{n,P} \rightsquigarrow_B Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{58}$$

The notation $\rightsquigarrow_B$ is defined in the Appendix and just means the usual notion of convergence the bootstrap law.

We derive the large sample distributions and validity of the weighted bootstrap for structural functionals via the functional delta method, which we modify to handle uniformity with respect to the underlying dgp P. We shall need the following assumption on the structural functionals.

Assumption 5 (Uniformly Continuous Hadamard Differentiability of Structural Functionals)**.** Suppose that for each $P \in \mathcal{P}$, $\rho = \rho_P$ is an element of a compact subset $\mathbb{D}_\rho \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}$. Suppose $\varrho \mapsto \phi(\varrho)$, a functional of interest, mapping $\mathbb{D}_\rho$ to $\ell^\infty(\mathcal{W})$, is Hadamard differentiable in ϱ with with derivative ϕ'_ρ, tangentially to $\mathbb{D}_0 = UC(\mathcal{U})^{d_\rho}$, uniformly in $\rho \in \mathbb{D}_\rho$, and that the mapping $(\varrho, h) \mapsto \phi'_\varrho(h)$ from $\mathbb{D}_\rho \times \mathbb{D}_0$ into $\ell^\infty(\mathcal{W})$ is defined and continuous. ■

This assumption holds for all examples of structural parameters we listed in Section 2.

The following result gives asymptotic Gaussian laws for $\sqrt{n}(\widehat{\Delta} - \Delta)$, the properly normalized structural estimates $\sqrt{n}(\widehat{\Delta} - \Delta)$. It also show that the bootstrap law of

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}),$$

computed conditionally on the data, approaches the asymptotic Gaussian law for $\sqrt{n}(\widehat{\Delta} - \Delta)$.

Theorem 4.3 (**Limit theory and Validity of Weighted Bootstrap For Smooth Structural Functionals**). *Under Assumptions 2, 3 or 4, and 5,*

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_\rho(Z_P), \quad \text{in } \ell^\infty(\mathcal{W})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{59}$$

where T_P is a zero mean tight Gaussian process, for each $P \in \mathcal{P}$. Moreover,

$$\sqrt{n}(\widehat{\Delta}^* - \widehat{\Delta}) \rightsquigarrow_B T_P := \phi'_\rho(Z_P), \quad \text{in } \ell^\infty(\mathcal{W})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{60}$$

5. Generic Lasso and Post-Lasso Methods for Functional Response Data

In this section, we provide estimation and inference results for Lasso and Post-Lasso estimators with function-valued outcomes and linear or logistic links. These results are of interest beyond the context of treatment effects estimation, and thus we present this section in a way that leaves it autonomous with respect to the rest of the paper.

5.1. **The generic setting with function-valued outcomes.** Consider a data generating process with a functional response variable $(Y_u)_{u\in\mathcal{U}}$ and observable covariates X satisfying for each $u \in \mathcal{U}$

$$\mathrm{E}[Y_u \mid X] = \Lambda(f(X)'\theta_u) + r_u(X), \tag{61}$$

where $f : \mathcal{X} \to \mathbb{R}^p$ is a set of p measurable transformations of the initial controls X, θ_u is a p-dimensional vector, r_u is an approximation error, and Λ is a fixed link function. We note that the notation in this section differs from the rest of the paper with Y_u and X denoting a generic response and generic covariates to facilitate the application of these results in other contexts. We only consider the cases of linear link function, $\Lambda(t) = t$, and the logistic link function $\Lambda(t) = \exp(t)/\{1 + \exp(t)\}$,[10] in detail; but we note that the principles discussed here apply to any M-estimator. In the remainder of the section, we discuss and establish results for ℓ_1-penalized and post-model selection estimators for $\theta_u, u \in \mathcal{U}$, that hold uniformly over $u \in \mathcal{U}$.

Throughout the section, we assume that $u \in \mathcal{U} \subset [0,1]^{d_u}$ and that i.i.d. observations from a dgp where (61) holds, $\{(Y_{ui}, u \in \mathcal{U}, X_i, f(X_i)) : i = 1, \ldots, n\}$, are available to estimate $(\theta_u)_{u\in\mathcal{U}}$. For $u \in \mathcal{U}$, a penalty level λ, and a diagonal matrix of penalty loadings $\widehat{\Psi}_u$, we define the Lasso estimator as

$$\widehat{\theta}_u \in \arg\min_\theta \mathbb{E}_n[M(Y_u, f(X)'\theta)] + \frac{\lambda}{n}\|\widehat{\Psi}_u\theta\|_1 \tag{62}$$

where $M(y,t) = \frac{1}{2}(y - \Lambda(t))^2$ for the case of linear regression, and $M(y,t) = 1(y = 1)\log\Lambda(t) + 1(y = 0)\log(1 - \Lambda(t))$ in the case of the logistic link function for binary response data. The corresponding Post-Lasso estimator is then defined as

$$\widetilde{\theta}_u \in \arg\min_\theta \mathbb{E}_n[M(Y_u, f(X)'\theta)] \quad : \quad \mathrm{supp}(\theta) \subseteq \mathrm{supp}(\widehat{\theta}_u). \tag{63}$$

The chief departure between analysis of Lasso and Post-Lasso when $\mathcal{U}$ is a singleton and the functional response case is that the penalty parameter needs to be set to control selection errors uniformly over $u \in \mathcal{U}$. To uniformly control these errors, we will set the penalty parameter λ so that with high probability

$$\frac{\lambda}{n} \geqslant c \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_u^{-1}\mathbb{E}_n[\nabla M(Y_u, f(X)'\theta_u)]\|_\infty. \tag{64}$$

[10]Considering the logistic link is useful for binary response data where $Y_u \in \{0,1\}$ for each $u \in \mathcal{U}$, though the linear link can be used in this case as well.

Similarly to Bickel, Ritov, and Tsybakov (2009); Belloni and Chernozhukov (2013); and Belloni, Chernozhukov, and Wang (2011) who use an analog of (64) appropriate for when $\mathcal{U}$ is a singleton in deriving properties Lasso and Post-Lasso estimators, guaranteeing that the "regularization event" (64) holds with high probability plays a key role in establishing desirable properties of Lasso and Post-Lasso estimators in the functional outcome case.

To implement (64), we propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}), \tag{65}$$

where d is the dimension of $\mathcal{U}$, $1-\gamma$ with $\gamma = o(1)$ is a confidence level associated with the probability of event (64), and $c > 1$ is a slack constant similar to that of Bickel, Ritov, and Tsybakov (2009). In practice, we set $c = 1.1$ and $\gamma = .1/\log(n)$ though other choices are theoretically valid.

In addition to the penalty parameter λ, we also need to construct a penalty loading matrix $\widehat{\Psi}_u = \mathrm{diag}(\{\widehat{l}_{uj,k}, j = 1,\ldots,p\})$. This loading matrix can be formed according to the following iterative algorithm.

Algorithm 1 (Estimation of Penalty Loadings)**.** Choose $\gamma \in (1/n, 1/\log n)$ can $c > 1$ to form λ as defined in (65), and choose a constant $K \geqslant 1$ as an upper bound on the number of iterations. (0) Set $k = 0$, and initialize $\widehat{l}_{uj,0}$ for each $j = 1,\ldots,p$. For the linear link function, set $\widehat{l}_{uj,0} = \{\mathbb{E}_n[f_j^2(X)(Y_u - \bar{Y}_u)^2]\}^{1/2}$ with $\bar{Y}_u = \mathbb{E}_n[Y_u]$. For the logistic link function, set $\widehat{l}_{uj,0} = \frac{1}{2}\{\mathbb{E}_n[f_j^2(X)]\}^{1/2}$. (1) Compute the Lasso and Post-Lasso estimators, $\widehat{\theta}_u$ and $\widetilde{\theta}_u$, based on $\widehat{\Psi}_u = \mathrm{diag}(\{\widehat{l}_{uj,k}, j = 1,\ldots,p\})$. (2) Set $\widehat{l}_{uj,k+1} := \{\mathbb{E}_n[f_j^2(X)(Y_u - \Lambda(f(X)'\widetilde{\theta}_u))^2]\}^{1/2}$. (3) If $k > K$, stop; otherwise set $k \leftarrow k+1$ and go to step (1).

5.2. **Asymptotic Properties of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses: Linear Case.** In the following, we provide sufficient conditions for establishing good performance of the estimators discussed in Section 5.1 when the linear link function is used. In the statement of the following assumption, $\delta_n \searrow 0$, $\ell_n \nearrow \infty$, and $\Delta_n \searrow 0$ are fixed sequences; and c, C, κ', κ'' and $\nu \in (0,1]$ are positive finite constants.

Assumption 6. For each $n \geqslant 1$, our data consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X)$ defined on the probability space $(S, \mathcal{S}, P)$ such that model (61) holds with $\mathcal{U} \subset [0,1]^{d_u}$. Consider $\Lambda(t) = t$ and $\zeta_u = Y_u - \mathrm{E}[Y_u \mid X]$. Suppose the following conditions hold uniformly for all $P \in \mathcal{P}_n$: (i) the model (61) is approximately sparse with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restriction $\log(pn/\gamma) \leqslant \delta_n n^{1/3}$. (ii) The set $\mathcal{U}$ has covering entropy bounded as $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant d\log(1/\epsilon) \vee 0$, and the collection $(Y_u, \zeta_u, r_u)_{u\in\mathcal{U}}$ is suitably measurable. (iii) Uniformly over $u \in \mathcal{U}$, the model's moments are boundedly heteroscedastic, namely $c \leqslant \mathrm{E}_P[\zeta_u^2 \mid X] \leqslant C$ and $\max_{j\leqslant p}\mathrm{E}_P[|f_j(X)\zeta_u|^3 + |f_j(X)Y_u|^3] \leqslant C$. (iv) We have that the dictionary functions, approximation errors, and empirical errors obey the following boundedness and empirical regularity conditions: (a) $c \leqslant \mathrm{E}_P[f_j^2(X)] \leqslant C$, $j = 1,\ldots,p$; $\max_{j\leqslant p}|f_j(X)| \leqslant K_n$ a.s.; $K_n \log(p\vee n) \leqslant \delta_n n^{\{\nu\wedge\frac{1}{2}\}}$. (b) With probability $1-\Delta_n$, $\sup_{u\in\mathcal{U}}\mathbb{E}_n[r_u^2(X)] \leqslant Cs\log(p\vee n)/n$;

$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \vee |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)Y_u^2]| \leqslant \delta_n$; $\sup_{d_{\mathcal{U}}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n + \mathrm{E}_P)[(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leqslant C\{\epsilon^\nu + n^{-1/2}\}$. (c) The empirical minimum and maximum sparse eigenvalues are bounded from zero and above, $\kappa' \leqslant \inf_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \sup_{\|\delta\|_0\leqslant s\ell_n} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leqslant \kappa''$. ■

Under Assumption 6, we establish results on the performance of the estimators (62) and (63) for the linear link function case that hold uniformly over $u \in \mathcal{U}$.

Theorem 5.1 (Rates and Sparsity for Functional Responses under Linear Link)**.** *Under Assumption 6 and setting penalties as in Algorithm 1, for all n large enough, uniformly for all $P \in \mathcal{P}_n$ with P_P probability $1-o(1)$, the Lasso estimator $\widehat{\theta}_u$ is uniformly sparse, $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u\|_0 \leqslant \bar{C}s$, and the following performance bounds hold for some constant $\bar{C}$:*

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \bar{C}\sqrt{\frac{s\log(p\vee n)}{n}} \text{ and } \sup_{u\in\mathcal{U}} \|\widehat{\theta}_u - \theta_u\|_1 \leqslant \bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

For all n large enough, uniformly for all $P \in \mathcal{P}_n$, with P_P probability $1-\delta-o(1)$, the Post-Lasso estimator corresponding to $\widehat{\theta}_u$ obeys

$$\sup_{u\in\mathcal{U}} \|f(X)'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \frac{\bar{C}}{\sqrt{\delta}}\sqrt{\frac{s\log(p\vee n)}{n}}, \text{ and } \sup_{u\in\mathcal{U}} \|\widetilde{\theta}_u - \theta_u\|_1 \leqslant \frac{\bar{C}}{\sqrt{\delta}}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

We note that the performance bounds are exactly of the type used in Assumptions 3 and 4.

5.3. **Asymptotic Properties of a Continuum of Lasso and Post-Lasso Estimators for Functional Responses: Logistic Case.** Next we provide sufficient conditions to state results on the performance of the estimators discussed above for the logistic link function. This case corresponds to $M(y,t) = 1(y=1)\log\Lambda(t) + 1(y=0)\log(1-\Lambda(t))$ with $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$ where the response variable is assumed to be binary, $Y_u \in \{0,1\}$ for all $u \in \mathcal{U}$.

Consider fixed sequences $\delta_n \to 0$, $\ell_n \nearrow \infty$, $\Delta_n \to 0$ and positive finite constants c, C, κ', κ'' and $\nu \in (0,1]$.

Assumption 7. For each $n \geqslant 1$, our data consist of i.i.d. copies $(W_i)_{i=1}^n$ of the stochastic process $W = ((Y_u)_{u\in\mathcal{U}}, X)$ defined on the probability space $(S, \mathcal{S}, P)$ such that model (61) holds with $\mathcal{U} \subset [0,1]^{d_u}$. Consider $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$, $Y_u \in \{0,1\}$, and $\zeta_u = Y_u - \mathrm{E}[Y_u \mid X]$. Suppose the following conditions hold uniformly for all $P \in \mathcal{P}_n$: (i) the model (61) is approximately sparse form with sparsity index obeying $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leqslant s$ and the growth restrictions $\log(pn/\gamma) \leqslant \delta_n n^{1/3}$ and $s\log(pn/\gamma) \leqslant \delta_n n$. (ii) The set $\mathcal{U}$ has covering entropy bounded as $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant d\log(1/\epsilon) \vee 0$, and the collection $(Y_u, \zeta_u, r_u)_{u\in\mathcal{U}}$ is suitably measurable. (iii) Uniformly over $u \in \mathcal{U}$ the model's moments are boundedly heteroscedastic, namely $c \leqslant \mathrm{E}_P[\zeta_u^2 \mid X] \leqslant C$, $\max_{j\leqslant p}\mathrm{E}_P[|f_j(X)\zeta_u|^3] \leqslant C$, and $\underline{c} \leqslant \mathrm{E}_P[Y_u \mid X] \leqslant 1 - \underline{c}$. (iv) We have that the dictionary functions, approximation errors, and empirical errors obey the following boundedness and empirical regularity conditions: (a) $\sup_{u\in\mathcal{U}} |r_u(X)| \leqslant \delta_n$ a.s.; $c \leqslant \mathrm{E}_P[f_j^2(X)] \leqslant C$, $j = 1,\ldots,p$; $\max_{j\leqslant p}|f_j(X)| \leqslant K_n$ a.s.; $K_n\log(p\vee n) \leqslant \delta_n n^{\{\nu\wedge\frac{1}{2}\}}$. (b) With probability $1-\Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leqslant Cs\log(p\vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \leqslant \delta_n$;

$\sup_{d_{\mathcal{U}}(u,u')\leqslant\epsilon}\{(\mathbb{E}_n+\mathrm{E}_P)[(\zeta_u-\zeta_{u'})^2]\}^{1/2}\leqslant C\{\epsilon^\nu+n^{-1/2}\}$. (c) The empirical minimum and maximum sparse eigenvalues are bounded from zero and above: $\kappa'\leqslant\inf_{\|\delta\|_0\leqslant s\ell_n}\|f(X)'\delta\|_{\mathbb{P}_n,2}\leqslant\sup_{\|\delta\|_0\leqslant s\ell_n}\|f(X)'\delta\|_{\mathbb{P}_n,2}\leqslant\kappa''$. ∎

The following result characterizes the performance of the estimators (62) and (63) for the logistic link function case under Assumption 7.

Theorem 5.2 (Rates and Sparsity for Functional Response under Logistic Link)**.** *Under Assumption 7 and setting penalties as in Algorithm 1, for all n large enough, uniformly for all $P\in\mathcal{P}_n$ with P_P probability $1-o(1)$, the following performance bounds hold for some constant $\bar{C}$:*

$$\sup_{u\in\mathcal{U}}\|f(X)'(\widehat{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leqslant\bar{C}\sqrt{\frac{s\log(p\vee n)}{n}}\ \text{ and }\ \sup_{u\in\mathcal{U}}\|\widehat{\theta}_u-\theta_u\|_1\leqslant\bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

Moreover, provided that $K_n^2s^2\log(p\vee n)\leqslant\delta_n n$, the estimator is uniformly sparse: $\sup_{u\in\mathcal{U}}\|\widehat{\theta}_u\|_0\leqslant\bar{C}s$. For all n large enough, uniformly for all $P\in\mathcal{P}_n$, with P_P probability $1-\delta-o(1)$, the Post-Lasso estimator corresponding to $\widehat{\theta}_u$ obeys

$$\sup_{u\in\mathcal{U}}\|f(X)'(\widetilde{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leqslant\frac{\bar{C}}{\sqrt{\delta}}\sqrt{\frac{s\log(p\vee n)}{n}},\ \text{ and }\ \sup_{u\in\mathcal{U}}\|\widetilde{\theta}_u-\theta_u\|_1\leqslant\frac{\bar{C}}{\sqrt{\delta}}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

We note that the performance bounds satisfy the conditions of Assumptions 3 and 4.

6. Estimating the Effect of 401(k) Participation on Financial Asset Holdings

As an illustration of the methods in this paper, we consider estimation of the effect of 401(k) participation on accumulated assets as in Abadie (2003) and Chernozhukov and Hansen (2004). The key problem in determining the effect of participation in 401(k) plans on accumulated assets is saver heterogeneity coupled with the fact that the decision of whether to enroll in a 401(k) is non-random. It is generally recognized that some people have a higher preference for saving than others. It also seems likely that those individuals with the highest unobserved preference for saving would be most likely to choose to participate in tax-advantaged retirement savings plans and would tend to have otherwise high amounts of accumulated assets. The presence of unobserved savings preferences with these properties then implies that conventional estimates that do not account for saver heterogeneity and endogeneity of participation will be biased upward, tending to overstate the savings effects of 401(k) participation.

To overcome the endogeneity of 401(k) participation, Abadie (2003) and Chernozhukov and Hansen (2004) adopt the strategy detailed in Poterba, Venti, and Wise (1994; 1995; 1996; 2001); and Benjamin (2003) who use data from the 1991 Survey of Income and Program Participation and argue that eligibility for enrolling in 401(k) plan in this data can be taken as exogenous after conditioning on a few observables of which the most important for their argument is income. The basic idea of their argument is that, at least around the time 401(k)'s initially became available, people were unlikely to be basing their employment decisions on whether an employer

offered a 401(k) but would instead focus on income. Thus, eligibility for a 401(k) could be taken as exogenous conditional on income, and the causal effect of 401(k) elgibility could be directly estimated by appropriate comparison across eligible and ineligible individuals.[11] Abadie (2003) and Chernozhukov and Hansen (2004) use this argument for the exogeneity of eligibity conditional on controls to argue that 401(k) eligibity provides a valid instrument for 401(k) participation and employ IV methods to estimate the effect of 401(k) participation on accumulated assets.

As a complement to the work cited above, we estimate various treatment effects of 401(k) participation on holdings of financial assets using high-dimensional methods. A key component of the argument underlying the exogeneity of 401(k) elibility is that eligibility may only be taken as exogenous after conditioning on income. Both Abadie (2003) and Chernozhukov and Hansen (2004) adopt this argument but control only for a small number of terms. One might wonder whether the small number of terms considered is sufficient to adequately control for income and other related confounds. At the same time, power to learn anything about the effect of 401(k) participation decreases as one controls more flexibly for confounds. The methods developed in this paper offer one resolution to this tension by allowing us to consider a very broad set of controls and functional forms under the assumption that among the set of variables we consider there is a relatively low-dimensional set that adequately captures the effect of confounds. This approach is more general than that pursued in Chernozhukov and Hansen (2004) or Abadie (2003) which both implicitly assume that confounding effects can adequately be controlled for by a small number of variables chosen *ex ante* by the researcher.

We use the same data as Abadie (2003), Benjamin (2003), and Chernozhukov and Hansen (2004). The data consist of 9915 observations at the household level drawn from the 1991 SIPP. For our analysis, we use net total financial assets as our outcome variable, Y.[12] Our treatment variable, D, is an indicator for having positive 401(k) balances; and our instruments, Z, is an indicator for working at a firm that offers a 401(k) plan. The vector of controls, X, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home ownership indicator. Further details about the sample and variables used can be found in Chernozhukov and Hansen (2004).

We present results for four different sets of control variables. The first set of control variables uses the indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, a linear term for family size, five categories for age, four categories for education, and seven categories for income (Indicator specification). We use the same definitions of categories as in Chernozhukov and Hansen (2004) and note that

[11] Poterba, Venti, and Wise (1994; 1995; 1996; 2001); and Benjamin (2003) all focus on estimating the effect of 401(k) eligibility, the intention to treat parameter. Also note that their are arguments that eligibility should not be taken as exogenous given income; see, for example, Engen and Scholz (1996) and Engen and Gale (2000).

[12] Net total financial assets are defined as the sum of checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks and mutual funds less nonmortgage debt, IRA balances, and 401(k) balances.

this is identical to the specification in Chernozhukov and Hansen (2004) and Benjamin (2003). The second specification uses the indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, and b-splines with three, four, six, and eight knots for family size, education, age, and income, respectively (B-Spline specification). The third specification augments the Indicator specification with all two-way interactions between the variables from the Indicator specification, and the fourth specification augments the B-Spline specification with all two-way interactions of the variables from the B-Spline specification. The dimensions of the set of control variables are thus 20, 27, 167, and 300 for the Indicator, B-Spline, Indicator plus interactions, and B-Spline plus interactions specifications, respectively.

We report estimates of the LATE, LATT, and LQTE for each of the four sets of control variables.[13] Estimation of all of the treatment effects depends on first-stage estimation of reduced form functions as detailed in Section 3. We estimate reduced form quantities where a transformation of Y is the outcome using least squares when no model selection is used or Post-Lasso when selection is used. We estimate propensity scores by logistic regression when no model selection is used or Post-ℓ_1-penalized logistic regression when selection is used. Penalty levels are chosen by cross-validation.

Estimates of the LATE and LATT are given in Table 1. In this table, we provide point estimates for each of the four sets of controls with and without variable selection. We also report both analytic and weighted-bootstrapped standard errors. The bootstapped standard errors are based on 500 bootstrap replications and standard exponential weights. Looking first at the two sets of standard error estimates, we see that the bootstrap and analytic standard are quite similar and that one would not draw substantively different conclusions from one versus the other. It is interesting that the estimated LATE and LATT are very similar in seven of the eight sets of estimates reported, suggesting positive and significant effects of 401(k) participation on net financial assets. The one exception is in the B-Spline specification with interactions in which both the LATE and LATT point estimates are implausibly large with associated very large estimated standard errors. One might be concerned that the instability in this case is due to there being important nonlinearity that is missed by the simpler specifications or the step-function approximation provided by including income categorically in the Indicator specification. This concern is alleviated by noting that the point estimate and standard error based on this set of controls following variable selection are sensible and similar to the other estimates. The fact that estimates following variable selection are similar to the other estimates suggests the bulk of the reduced form predictive power is contained in a set of variables similar to those used in the other specifications and that there are not a small number of the added variables that pick out important sources of nonlinearity neglected by the other specifications. Thus, the large point estimates and standard errors in this case seem to be driven by including many variables which have little to no predictive power in the reduced form relationships but result in overfitting.

[13] We focus on this set of treatment effects for brevity as they are sufficient to illustrate the results.

We provide estimates of the LQTE based on the Indicator specification and the B-Spline specification in Figures 1 and 2, respectively. In each figure, the top rows correspond to the specification without interactions while the bottom rows include the full set of interactions. Treatment effect estimates without variable selection are given in the left panels of each figure, and the results based on variable selection are given in the right panels. We report point estimates as the solid line in each graphic and report uniform 95% confidence intervals with dashed lines. As with estimates of the LATE, we see that estimates of the LQTE based on variable selection methods are stable regardless of which set of variables we consider. This stability differs sharply from the behavior of the non-regularized estimators which provide erratic point estimates and very wide confidence intervals. Again, this erratic behavior is likely due to overfitting as the variable selection methods select a roughly common low-dimensional set of variables that are useful for reduced form prediction in all cases.

If we focus on the LQTE estimated from variable selection methods, we find that 401(k) participation has a small impact on accumulated financial assets at low quantiles while appearing to have a much larger impact at high quantiles. Looking at the uniform confidence intervals, we can see that this pattern is statistically significant at the 5% level and that we would reject a pattern of constant treatment effects. Interpreting the quantile index as "preference for savings" as in Chernozhukov and Hansen (2004), these results suggest that 401(k) participation has little causal impact on the accumulated financial assets of those with low desire to save but a much larger impact on those with stronger preferences for saving.

It is interesting that our results are quite similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of control variables. The similarity is due to the fact that the variable selection methods consistently pick a set of variables similar to those used in previous work. The fact that we allow for a rich set of controls but produce similar results to those previously available lends further credibility to the claim that previous work controlled adequately for the available observables.[14] Finally, it is worth noting that this similarity is not mechanical or otherwise built in to the procedure. For example, applications in Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni, Chernozhukov, and Hansen (2011) use high-dimensional variable selection methods and produce sets of variables that differ substantially from intuitive baselines.

Appendix A. Notations and Tools

A.1. **Stochastic Convergence Uniformly in P.** All parameters, such as the law of the data, are indexed by P, sometimes referred to as the the data-generating process. This dependency, which is well understood, is kept implicit throughout. We shall allow the possibility that the probability measure $P = P_n$ can depend on n. We shall conduct our stochastic convergence analysis uniformly in P, where P can vary within some set $\mathcal{P}_n$, which itself may vary with

[14]Of course, the estimates are still not valid causal estimates if one does not believe that 401(k) eligibiliy can be taken as exogenous after controlling for income and the other included variables.

n. The convergence analysis, namely stochastic order relations and convergence in distribution, uniformly in $P \in \mathcal{P}_n$ and the analysis under all sequences $P_n \in \mathcal{P}_n$ are equivalent. Specifically, consider the metric space $\ell^\infty(\mathcal{U})$, the space of uniformly bounded functions mapping an arbitrary index set $\mathcal{U}$ to the real line.

Consider in this space a sequence of stochastic processes X_n and a random element Y, taking values in $\mathbb{D}$, defined on the probability space $(\Omega, \mathcal{A}, \mathrm{P}_P)$. Also consider the sequence of stochastic processes X_n and a fixed random element Y , taking values in $\ell^\infty(\mathcal{U})$, and defined on the probability space $(\Omega, \mathcal{A}, \mathrm{P}_P)$. Consider a sequence of deterministic positive constants a_n. We shall say that

(i) $X_n = O_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\lim_{K \nearrow \infty} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}^*_P(|X_n| > Ka_n) = 0$,
(ii) $X_n = o_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{K>0} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}^*_P(|X_n| > Ka_n) = 0$,
(iii) $X_n \rightsquigarrow Y$ uniformly in $P \in \mathcal{P}_n$, if

$$\sup_{P\in\mathcal{P}_n} \sup_{h\in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}^*_{\mathrm{P}} h(X_n) - \mathrm{E}_{\mathrm{P}} h(Y)| \to 0$$

.

Here the error $\rightsquigarrow$ denotes weak convergence, i.e. convergence in distribution or law. The symbol $\rho_{w,P}(X, Y)$ denotes the bounded Lipschitz metric that measures distance between the law of X and the of law Y, where outer expectations are used in cases X_n is not measurable. This is a distance that metrizes weak convergence, with outer expectations are used to handle measurability issues; see VW.

Lemma A.1. *The above notions are equivalent to the following notions:*

(i) *for every sequence* $P_n \in \mathcal{P}_n$, $X_n = O_{P_n}(a_n)$, *i.e.* $\lim_{K\nearrow\infty} \lim_{n\to\infty} \mathrm{P}^*_{P_n}(|X_n| > Ka_n) = 0$,
(ii) *for every sequence* $P_n \in \mathcal{P}_n$, $X_n = o_{P_n}(a_n)$, *i.e.* $\sup_{K>0} \lim_{n\to\infty} \mathrm{P}^*_n(|X_n| > Ka_n) = 0$,
(iii) *for every sequence* $P_n \in \mathcal{P}_n$, $X_n \rightsquigarrow Y$, *i.e.*

$$\sup_{h\in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}^*_{\mathrm{P}_n} h(X_n) - \mathrm{E}_{\mathrm{P}_n} h(Y)| \to 0$$

.

Proof of Lemma A.1. The claims follow straightforwardly from definitions, and so the proof is omitted.

A.2. **Uniform Donsker Property and Uniform Pre-Gaussianity.** We shall invoke the following lemma.

Lemma A.2. *Let* $\mathcal{F} : S \mapsto \mathbb{R}$ *be an image-admissible Suslin class of functions with a measurable envelope* $F : S \mapsto \mathbb{R}$. *Furthermore, suppose that*

$$\lim_{M\nearrow\infty} PF^2 1\{F > M\} = 0, \quad \int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L^2(Q))} d\epsilon < \infty.$$

Then the class $\mathcal{F}$ is Donsker uniformly in $P \in \mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \sup_{h\in BL_1} |\mathrm{E}_P^* h(\mathbb{G}_{n,P}) - \mathrm{E}_P h(\mathbb{G}_P)| \to 0.$$

Moreover, $\mathcal{F}$ is pre-Gaussian uniformly in $P \in \mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{f\in\mathcal{F}} |\mathbb{G}_P(f)| < \infty, \quad \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{\|f-g\|_{P,2}\leqslant\delta} |\mathbb{G}_P(f) - \mathbb{G}_P(g)| = 0.$$

Proof. This is an immediate consequence of Theorem 2.8.2 in van der Vaart and Wellner (1996). The image admissible Suslin condition, defined on page 186 in Dudley (2000), implies the measurability conditions needed in Theorem 2.8.2 in van der Vaart and Wellner (1996), e.g. by a reasoning given in Example 2.3.5. van der Vaart and Wellner (1996) ∎

A.3. **Probabilistic Inequalities.** Let $N(\epsilon, \mathcal{F}, \|\cdot\|_{Q,2})$ denote the ϵ-covering number of $\mathcal{F}$ with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$, where Q is finitely discrete. Let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} \mathrm{P} f^2 \leqslant \sigma^2 \leqslant \|F\|_{\mathrm{P},2}^2$. Let $M = \max_{1\leqslant i\leqslant n} F(X_i)$.

Lemma A.3 (**A Maximal Inequality**). *Let $\mathcal{F}$ be an image admissible Suslin set set of functions with a measurable envelope F. Suppose that $F = \sup_{f\in\mathcal{F}} |f| \in \mathcal{L}^q(P)$ for some $q \geqslant 2$. Let $M = \max_{i\leqslant n} F(W_i)$. Suppose that there exist constants $a \geqslant e$ and $v \geqslant 1$ such that*

$$\log \sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leqslant v(\log a + \log(1/\varepsilon)), \ 0 < \forall\varepsilon \leqslant 1.$$

Then

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim \sqrt{v\sigma^2 \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P},2}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right).$$

Moreover, for every $t \geqslant 1$, with probability $> 1 - t^{-q/2}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (1+\alpha)\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q)\Big[(\sigma + n^{-1/2}\|M\|_{\mathrm{P},q})\sqrt{t} + \alpha^{-1}n^{-1/2}\|M\|_{\mathrm{P},2}t\Big], \ \forall\alpha > 0,$$

where $K(q) > 0$ is a constant depending only on q.

Proof. See Chernozhukov, Chetverikov, Kato (2012). ∎

Lemma A.4 (**A Self-Normalized Maximal Inequality**). *Let $\mathcal{F}$ be an image-admissible Suslin set of functions with a measurable envelope F. Suppose that $F \geqslant \sup_{f\in\mathcal{F}} |f| \geqslant 1$, and suppose that there exist some constants $p > 1$, $m \geqslant 1$, and $\kappa \geqslant 3 \vee n$ such that*

$$\log N(\epsilon\|F\|_{\mathbb{P}_n,2}, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n,2}) \leqslant (\kappa/\epsilon)^m, \ 0 < \epsilon < 1.$$

Then for every $\delta \in (0, 1/6)$, with probability at least $1 - \delta$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (C'/\sqrt{\delta})\sqrt{m\log(\kappa\|F\|_{\mathbb{P}_n,2})} \max\left\{\sup_{f\in\mathcal{F}} \|f\|_{\mathrm{P},2}, \sup_{f\in\mathcal{F}} \|f\|_{\mathbb{P}_n,2}\right\},$$

where the constant C' is universal.

Proof. The inequality can be deduced from (Belloni and Chernozhukov, 2011b), with the exception that the envelope is allowed to be larger than $\sup_{f\in\mathcal{F}}|f|$. ■

Lemma A.5 (**Algebra for Covering Entropies**).

(1) *Let $\mathcal{F}$ be a measurable VC class with a finite VC index k or any other class whose entropy is bounded above by that of such a VC class, then the covering entropy of F obeys:*

$$\sup_Q \log N(\epsilon\|F\|_{Q,2},\mathcal{F},L^2(Q)) \lesssim 1+k\log(1/\epsilon)$$

Examples include $\mathcal{F}=\{\alpha' z,\alpha\in\mathbb{R}^k,\|\alpha\|\leqslant C\}$ and $\mathcal{F}=\{1\{\alpha' z>0\},\alpha\in\mathbb{R}^k,\|\alpha\|\leqslant C\}$.

(2) *For any measurable function sets $\mathcal{F}$ and $\mathcal{F}'$:*

$$\begin{aligned}
&\log N(\epsilon\|F+F'\|_{Q,2},\mathcal{F}+\mathcal{F}',L^2(Q)) \leqslant B\\
&\log N(\epsilon\|F\cdot F'\|_{Q,2},\mathcal{F}\cdot\mathcal{F}',L^2(Q)) \leqslant B\\
&\log N(\epsilon\|F\vee F'\|_{Q,2},\mathcal{F}\cup\mathcal{F}',L^2(Q)) \leqslant B\\
&B=\log N\left(\frac{\epsilon}{2}\|F\|_{Q,2},\mathcal{F},L^2(Q)\right)+\log N\left(\frac{\epsilon}{2}\|F'\|_{Q,2},\mathcal{F}',L^2(Q)\right).
\end{aligned}$$

(3) *Given a measurable class $\mathcal{F}$ and a random variable g_i:*

$$\log\sup_Q N(\epsilon\||g|F\|_{Q,2},g\mathcal{F},L^2(Q)) \lesssim \log\sup_Q N\left(\epsilon/2\|F\|_{Q,2},\mathcal{F},L^2(Q)\right)$$

(4) *For the class $\mathcal{F}^*$ created by integrating $\mathcal{F}$, i.e. $f^*(x):=\int f(x,y)d\mu(y)$ where μ is any probability measure,*

$$\log N(\epsilon\|F\|_{Q,2},\mathcal{F}^*,L^2(Q)) \leqslant \log N\left(\epsilon\|F\|_{Q,2},\mathcal{F},L^2(Q)\right)$$

Proof. For the proof of assertions (1)-(3) see, e.g., (Andrews, 1994). The fact (4) was noted in Chandraksekhar et al (2011), though it is rather elementary and follows from convexity of the norm and Jensen's inequality: $\|f^*-\tilde{f}^*\|_{Q,2}\leqslant\int\|f-\tilde{f}\|_{Q,2}d\mu=\|f-\tilde{f}\|_{Q,2}$, from which the stated bound follows immediately. In other words, any averaging done over components of the function contracts distances between functions and therefore does not expand the covering entropy. A related, slightly different bound is stated in Ghosal and Van der Vaart (2009), but we need the bound above. ■

Lemma A.6 (**Contractivity of Conditional Expectation**). *Let (V,X) and (V',X) be random vectors in $\mathbb{R}\times\mathbb{R}^k$ defined on the probability space $(S,\mathcal{S},Q)$, with the first components being scalar, then for any $1\leqslant q\leqslant\infty$,*

$$\|\mathrm{E}_Q(V|X)-\mathrm{E}_Q(V'|X)\|_{Q,q}\leqslant\|V-V'\|_{Q,q}.$$

This is an instance of a well known result on the contractive property of the conditional expectation. We recall it here since we shall use if frequently.

A.4. **Hadamard Differentiability for Sequences and Delta Method for Sequences.** We shall use the functional delta method, as formulated in VW. Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces, with $\mathbb{D}_0 \subset \mathbb{D}$. A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is called *Hadamard-differentiable* at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ if there is a continuous linear map $\phi'_\theta : \mathbb{D}_0 \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h), \quad n \to \infty,$$

for all sequences $t_n \to 0$ and $h_n \to h \in \mathbb{D}_0$ such that $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n.

A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is called *Hadamard-differentiable uniformly* in $\theta \in \mathbb{D}_\phi$, a compact subset of $\mathbb{D}$, tangentially to $\mathbb{D}_0$, if

$$\left| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right| \to 0, \quad n \to \infty,$$

for all sequence $\theta_n \to \theta$ and for all sequences $t_n \to 0$ and $h_n \to h \in \mathbb{D}_0$ such that $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n. As a part of the definition, we require that the map $h \mapsto \phi'_\theta(h)$ from $\mathbb{D}_0$ to $\mathbb{E}$ is continuous and linear, and that the map $(\theta, h) \mapsto \phi'_\theta(h)$ from $\mathbb{D}_\phi \times \mathbb{D}_0$ to $\mathbb{E}$ is continuous.

Lemma A.7 (Functional delta-method for sequences)**.** *Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$. Let X_n be a sub-sequence of stochastic processes taking values in $\mathbb{D}_\phi$ such that $r_n(X_n - \theta_n) \rightsquigarrow X$ and $\theta_n \to \theta$ in $\mathbb{D}$ along a subsequence $n \in \mathbb{Z}' \subset \mathbb{Z}$, where X possibly depends on $\mathbb{Z}'$ and is separable and takes its values in $\mathbb{D}_0$, for some sequence of constants $r_n \to \infty$. Then $r_n(\phi(X_n) - \phi(\theta_n)) \rightsquigarrow \phi'_\theta(X)$ in $\mathbb{E}$ along the same subsequence. If ϕ'_θ is defined and continuous on the whole of $\mathbb{D}$, then the sequence $r_n(\phi(X_n) - \phi(\theta_n)) - \phi'_{\theta_n}(r_n(X_n - \theta_n))$ and $\phi'_{\theta_n}(r_n(X_n - \theta_n)) - \phi'_{\theta_n}(r_n(X_n - \theta_n))$ converges to zero in outer probability along the same subsequence.*

Let $D_n = (W_i)_{i=1}^n$ denote the data vector and $M_n = (\xi_i)_{i=1}^n$ be a vector of random variables, used to generate bootstrap draws or simulation draws (this may depend on particular method). Consider sequences of stochastic processes $V_n(D_n)$, where the sequence $G_n = \sqrt{n}(V_n - V)$ weakly converges unconditionally to the tight random element G. This means that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}^*_{P_n} h(G_n) - \mathrm{E}_G h(G)| \to 0,$$

along $n \in \mathbb{Z}'$, where E_G denotes the expectation computed with respect to the law of G. This is denoted as denoted as $G_n \rightsquigarrow G$ along $n \in \mathbb{Z}'$. Also consider the bootstrap stochastic process $G^*_n = G_n(D_n, M_n)$ in a normed space $\mathbb{D}$, where G_n is a measurable function of M_n for each value of D_n. Suppose that G^*_n converges conditionally given D_n in distribution to G, in probability, that is

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D},\mathbb{R})} |\mathrm{E}_{M_n}[h(G^*_n)] - \mathrm{E}_G h(G)| \to 0,$$

in outer probability along $n \in \mathbb{Z}'$, where E_{M_n} denotes the expectation computed with respect to the law of M_n holding the data D_n fixed. This is denoted as $G^*_n \rightsquigarrow_B G$ along $n \in \mathbb{Z}'$, respectively.

Let $V^*_n = V_n + G^*_n/\sqrt{n}$ denote the bootstrap or simulation draw of V_n.

Lemma A.8 (Delta-method for bootstrap and other simulation methods)**.** *Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces, with $\mathbb{D}_0 \subset \mathbb{D}$. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at V tangentially to $\mathbb{D}_0$, with the derivative map ϕ'_V. Let V_n and V_n^* be maps as indicated previously with values in $\mathbb{D}_\phi$ such that $\sqrt{n}(V_n - V) \rightsquigarrow G$ and $\sqrt{n}(V_n^* - V_n) \rightsquigarrow_B G$ in $\mathbb{D}$ along a subsequence of integers $n \in \mathbb{Z}' \subset \mathbb{Z}$, where G is separable and takes its values in $\mathbb{D}_0$. Then $\sqrt{n}(\phi(V_n^*) - \phi(V_n)) \rightsquigarrow_B \phi'_V(G)$ in $\mathbb{E}$ along the same subsequence.*

Another technical result that we use in the sequel concerns the equivalence of continuous and uniform convergence.

Lemma A.9 (Uniform convergence via continuous convergence)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be complete separable metric spaces, with $\mathbb{D}$ compact. Suppose $f : \mathbb{D} \mapsto \mathbb{E}$ is continuous. Then a sequence of functions $f_n : \mathbb{D} \mapsto \mathbb{E}$ converges to f uniformly on $\mathbb{D}$ if and only if for any convergent sequence $x_n \to x$ in $\mathbb{D}$ we have that $f_n(x_n) \to f(x)$.*

Proofs of Lemmas A.7 and A.8. The result follows from the proofs in VW, Chap. 3.9, where proofs (pointwise in θ) are given for sequence of integers $n \in \{1, 2, ...\}$. The claim extends to subsequences trivially. In the proof of Lemma A.7, the extension to $\theta_n \to \theta$ case, along subsequences, follows by combining their arguments with Lemma A.9 ■

Appendix B. Proofs for Section 4

B.1. **Proof of Theorem 4.1.** The two results for the two strategies have similar structure, so we only give the proof for one of the strategies – the first strategy.

Step 0. (A Preamble). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions only, but not on n once $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$, and not on $\mathrm{P} \in \mathcal{P}_n$. We consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $\mathrm{P} = \mathrm{P}_n$ throughout the proof, *suppressing* the index n. Since the argument is asymptotic, we can just assume that $n \geqslant n_0$ in what follows.

To proceed with presentation of proofs, it might be convenient for the reader to have notations collected in one place. The influence function and low-bias moment functions for $\alpha_V(z)$ for $z \in \mathcal{Z} = \{0, 1\}$ are given respectively by:

$$\psi^\alpha_{V,z}(W) := \psi^\alpha_{V,z,g_V,m_Z}(W, \alpha_V(z)), \quad \psi^\alpha_{V,z,g,m}(W, \alpha) := \frac{1(Z = z)(V - g(z, X))}{m(z, X)} + g(z, X) - \alpha.$$

The influence functions and the moment functions for γ_V are given by $\psi^\gamma_V(W) := \psi^\gamma_V(W, \gamma_V)$ and $\psi^\gamma_V(W, \gamma) := V - \gamma$. Recall that the the estimator of the reduced-form parameters $\alpha_V(z)$ and $\gamma_V(z)$ are solutions $\alpha = \widehat{\alpha}_V(z)$ and $\gamma = \widehat{\gamma}_V$ to the equations:

$$\mathbb{E}_n[\psi^\alpha_{V,z,\widehat{g}_V,\widehat{m}_Z}(W, \alpha)] = 0, \quad \mathbb{E}_n[\psi^\gamma_V(W, \gamma)] = 0,$$

where $\widehat{g}_V(z, x) = \Lambda_V(f(z, x)'\bar{\beta}_V)$ and $\widehat{m}_Z(z, x) = \Lambda_Z(f(z, x)'\bar{\beta}_Z)$, where $\bar{\beta}_V$ and $\bar{\beta}_Z$ are as in Asssumption 3. For each variable name $V \in V_u$,

$$V_u := (V_{uj})_{j=1}^5 := (Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)Y_u),$$

we obtain the estimator $\widehat{\rho}_u := \left(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1), \widehat{\gamma}_V\}\right)_{V \in V_u}$ of $\rho_u := \left(\{\alpha_V(0), \alpha_V(1), \gamma_V\}\right)_{V \in V_u}$. The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with a finite dimension. We stack these vectors into processes $\rho = (\rho_u)_{u \in \mathcal{U}}$ and $\widehat{\rho} = (\widehat{\rho}_u)_{u \in \mathcal{U}}$.

Step 1.(Linearization) In this step we establish the first claim, namely that

$$\sqrt{n}(\widehat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \tag{66}$$

where $Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}}$. The components $(\sqrt{n}(\widehat{\gamma}_{V_{uj}} - \gamma_{V_{uj}}))_{u \in \mathcal{U}}$ of $\sqrt{n}(\widehat{\rho} - \rho)$ trivially have the linear representation (with no error) for each $j \in \mathcal{J}$. We only need to establish the claim for the empirical process $(\sqrt{n}(\widehat{\alpha}_{V_{uj}}(z) - \alpha_{V_{uj}}(z)))_{u \in \mathcal{U}}$ for $z \in \{0, 1\}$, which we do in the steps below.

(a) We make some preliminary observations. For $t = (t_1, t_2, t_3, t_4) \in \mathbb{R}^2 \times (0, 1)^2$ and $v \in \mathbb{R}$, $(z, \bar{z}) \in \{0, 1\}^2$, we define the function $(v, z, \bar{z}, t) \mapsto \psi(v, z, \bar{z}, t)$ via:

$$\psi(v, z, 1, t) = \frac{1(z = 1)(v - t_2)}{t_4} + t_2, \quad \psi(v, z, 0, t) = \frac{1(z = 0)(v - t_1)}{t_3} + t_1.$$

The derivatives of this function with respect to t obey for all $k = (k_j)_{j=1}^4 \in \mathbb{N}^4 : 0 \leqslant |k| \leqslant 4$,

$$|\partial_t^k \psi(v, z, \bar{z}, t)| \leqslant L, \quad \forall (v, \bar{z}, z, t) : |v| \leqslant C, |t_1|, |t_2| \leqslant C, c'/2 \leqslant |t_3|, |t_4| \leqslant 1 - c'/2, \tag{67}$$

where L depends only on c' and C, $|k| = \sum_{j=1}^4 k_j$, and $\partial_t^k := \partial_{t_1}^{k_1} \partial_{t_2}^{k_2} \partial_{t_3}^{k_3} \partial_{t_4}^{k_4}$.

(b). Let

$$\begin{aligned}
&\widehat{h}_V(X_i) := (\widehat{g}_V(0, X_i), \widehat{g}_V(1, X_i), 1 - \widehat{m}(1, X_i), \widehat{m}(1, X_i),)', \\
&h_V(X_i) := (g_V(0, X_i), g_V(1, X_i), m(0, X_i), m(1, X_i))', \\
&f_{\widehat{h}_V, V, z}(W) := \psi(V, Z, z, \widehat{h}_V(X_i)), \\
&f_{h_V, V, z}(W) := \psi(V, Z, z, h_V(X_i)).
\end{aligned}$$

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

$$\widehat{g}_V(0, \cdot) \in \mathcal{G}_V(0), \quad \widehat{g}_V(1, \cdot) \in \mathcal{G}_V(1), \quad \widehat{m}(1, \cdot) \in \mathcal{M}(1), \quad \widehat{m}(0, \cdot) \in \mathcal{M}(0) = 1 - \mathcal{M}(1),$$

where

$$\mathcal{G}_V(d) := \left\{ \begin{array}{l} (x, z) \mapsto \Lambda_V(f(z, x)'\beta) : \|\beta\|_0 \leqslant sC \\ \|\Lambda_V(f(Z, X)'\beta) - g_V(d, Z, X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\ \|\Lambda_V(f(Z, X)'\beta) - g_V(d, Z, X)\|_{\mathrm{P},\infty} \lesssim \epsilon_n \end{array} \right\},$$

$$\mathcal{M}(1) := \left\{ \begin{array}{l} x \mapsto \Lambda_Z(f(x)'\beta) : \|\beta\|_0 \leqslant sC \\ \|\Lambda_Z(f(X)'\beta) - m_Z(1, X)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\ \|\Lambda_Z(f(X)'\beta) - m_Z(1, X)\|_{\mathrm{P},\infty} \lesssim \epsilon_n \end{array} \right\}.$$

To see this note, that under Assumption 3, under conditions (i)-(ii), under the event occurring under condition (ii) of that assumption: for all $n \geqslant \min\{j : \delta_j \leqslant 1/2\}$,

$$\begin{aligned}
\|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},2} &\leqslant \|\Lambda_Z(f(X)'\beta) - \Lambda_Z(f(X)'\beta_Z)\|_{\mathrm{P},2} + \|r_Z\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathrm{P},2} + \|r_Z\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathbb{P}_n,2} + \|r_Z\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\
\|\Lambda_Z(f(X)'\beta) - m_Z(1,X)\|_{\mathrm{P},\infty} &\leqslant \|\Lambda_Z(f(X)'\beta) - \Lambda_Z(f(X)'\beta_Z)\|_{\mathrm{P},\infty} + \|r_Z\|_{\mathrm{P},\infty} \\
&\leqslant \|\partial\Lambda_Z\|_\infty \|f(X)'(\beta-\beta_Z)\|_{\mathrm{P},\infty} + \|r_Z\|_{\mathrm{P},\infty} \\
&\lesssim K_n\|\beta-\beta_Z\|_1 + \epsilon_n \leqslant 2\epsilon_n,
\end{aligned}$$

for $\beta = \widehat{\beta}_Z$, with evaluation after computing the norms, and for $\|\partial\Lambda\|_\infty$ denoting $\sup_{l\in\mathbb{R}} |\partial\Lambda(l)|$ here and below. Similarly, under Assumption 3,

$$\begin{aligned}
\|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},2} &\lesssim \|\partial\Lambda_V\|_\infty \|f(Z,X)'(\beta-\beta_V)\|_{\mathbb{P}_n,2} + \|r_V\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\
\|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},\infty} &\lesssim K_n\|\beta-\beta_V\|_1 + \epsilon_n \leqslant 2\epsilon_n,
\end{aligned}$$

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

$$\|\Lambda_V(f(0,X)'\beta) - g_V(0,X)\|_{\mathrm{P},2} \vee \|\Lambda_V(f(1,X)'\beta) - g_V(1,X)\|_{\mathrm{P},2} \lesssim \|\Lambda_V(f(1,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},2}$$

under condition (iii) of Assumption 2, and trivially

$$\|\Lambda_V(f(0,X)'\beta) - g_V(0,X)\|_{\mathrm{P},\infty} \vee \|\Lambda_V(f(1,X)'\beta) - g_V(1,X)\|_{\mathrm{P},\infty} \leqslant \|\Lambda_V(f(Z,X)'\beta) - g_V(Z,X)\|_{\mathrm{P},\infty}$$

under condition (iii) of Assumption 2.

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

$$\widehat{h}_V \in \mathcal{H}_{V,n} := \{h = (\bar{g}(0,\cdot), \bar{g}(1,\cdot), \bar{m}(0,\cdot), \bar{m}(1,\cdot),) \in \mathcal{G}_V(0) \times \mathcal{G}_V(1) \times \mathcal{M}(0) \times \mathcal{M}(1)\}.$$

(c) We have that

$$\alpha_V(z) = \mathrm{E}[f_{h_V,V,z}] \text{ and } \widehat{\alpha}(z) = \mathbb{E}_n[f_{\widehat{h}_V,V,z}],$$

so that

$$\sqrt{n}(\widehat{\alpha}_V(z) - \alpha_V(z)) = \underbrace{\mathbb{G}_n[f_{h_V,V,z}]}_{I_V(z)} + \underbrace{(\mathbb{G}_n[f_{h,V,z}] - \mathbb{G}_n[f_{h_V,V,z}])}_{II_V(z)} + \underbrace{\sqrt{n}(\mathrm{E}[f_{h,V,z} - f_{h_V,h,z}])}_{III_V(z)},$$

with h evaluated at $h = \widehat{h}_V$.

(d) Note that for $\Delta_{V,i} = h(Z_i,X_i) - h_V(Z_i,X_i)$,

$$\begin{aligned}
III_V(z) &= \sqrt{n}\sum_{|k|=1} \mathrm{E}[\partial_t^k \psi(V_i, Z_i, z, h_V(Z_i,X_i))\Delta_{V,i}^k] \\
&+ \sqrt{n}\sum_{|k|=2} 2^{-1}\mathrm{E}[\partial_t^k \psi(V_i, Z_i, z, h_V(Z_i,X_i))\Delta_{V,i}^k] \\
&+ \sqrt{n}\sum_{|k|=3} \int_0^1 6^{-1}\mathrm{E}[\partial_t^k \psi(V_i, Z_i, z, h_V(Z_i,X_i) + \lambda\Delta_{V,i}^k)\Delta_{V,i}^k]d\lambda, \\
&=: III_V^a(z) + III_V^b(z) + III_V^b(z),
\end{aligned}$$

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

$$\mathrm{E}[\partial_t^k \psi(V_i, Z_i, z, h_V(Z_i, X_i))|Z_i, X_i] = 0 \quad \forall k \in \mathbb{N}^3 : |k| = 1,$$

we have that $III_V^a(z) = 0$. Moreover, uniformly for any $h \in \mathcal{H}_{V,n}$ we have that, in view of properties noted in step (a),

$$|III_V^b(z)| \lesssim \sqrt{n}\|h - h_V\|_{\mathrm{P},2}^2 \lesssim \sqrt{n}(\delta_n n^{-1/4})^2 \leqslant \delta_n^2,$$

$$|III_V^c(z)| \lesssim \sqrt{n}\|h - h_V\|_{\mathrm{P},2}^2 \|h - h_V\|_{\mathrm{P},\infty} \lesssim \sqrt{n}(\delta_n n^{-1/4})^2 \epsilon_n \leqslant \delta_n^2 \epsilon_n.$$

Since $\widehat{h}_V \in \mathcal{H}_{V,n}$ for all $V \in \mathcal{V}$ with probability $1 - \Delta_n$, we have that once $n \geqslant n_0$,

$$\mathrm{P}_P\Big(|III_V(z)| \lesssim \delta_n^2, \forall z \in \{0,1\}, \forall V \in \mathcal{V}\Big) \geqslant 1 - \Delta_n.$$

(e). Furthermore, we have that

$$\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |II_V(z)| \leqslant \sup_{h \in \mathcal{H}_{V,n}, z \in \{0,1\}, V \in \mathcal{V}} |\mathbb{G}_n[f_{h,V,z}] - \mathbb{G}_n[f_{h_V,V,z}]|.$$

The classes of functions, viewed as maps from the sample space S to the real line,

$$\mathcal{V} := \{V_{uj}, u \in \mathcal{U}, j \in \mathcal{J}\} \quad \text{and} \quad \mathcal{V}^* := \{g_{V_{uj}}(Z, X), u \in \mathcal{U}, j \in \mathcal{J}\}$$

are bounded by a constant envelope and have the uniform covering ϵ-entropy bounded by a multiple of $\log(\mathrm{e}/\epsilon) \vee 0$, that is $\log \sup_Q N(\varepsilon, \mathcal{V}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$, which holds by Assumption 2, and $\log \sup_Q N(\varepsilon, \mathcal{V}^*, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$ which holds by contractivity of conditions expectations noted in Lemma A.6 (or by Lemma A.5, item (iv)). The uniform covering ϵ-entropy of the function set $\mathcal{B} = \{1(Z = z), z \in \{0,1\}\}$ is trivially bounded by $\log(\mathrm{e}/\epsilon) \vee 0$.

The class of functions

$$\mathcal{G} := \{\mathcal{G}_V(d), V \in \mathcal{V}, d \in \{0,1\}\}$$

has a constant envelope and is a subset of

$$\{(x,z) \mapsto \Lambda(f(z,x)'\beta) : \|\beta\|_0 \leqslant sC, \Lambda \in \mathcal{L} = \{\mathrm{Id}, \Phi, 1 - \Phi, \Lambda_0, 1 - \Lambda_0\}\},$$

which is a union of 5 sets of the form

$$\{(x,z) \mapsto \Lambda(f(z,x)'\beta) : \|\beta\|_0 \leqslant sC\}$$

with $\Lambda \in \mathcal{L}$ a fixed monotone function for each of the 5 sets; each of these sets are the unions of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$ (note that a fixed monotone transformations Λ preserves the VC-subgraph property). Therefore

$$\log \sup_Q N(\varepsilon, \mathcal{G}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0.$$

Similarly, the class of functions $\mathcal{M} = (\mathcal{M}(1) \cup (1 - \mathcal{M}(1)))$ has a constant envelope, which is a union of at most 5 sets, which are themselves the unions of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$ (a fixed monotone transformations Λ preserves the VC-subgraph property). Therefore, $\log \sup_Q N(\varepsilon, \mathcal{M}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0$.

Finally, the set of functions

$$\mathcal{F}_n = (f_{h,V,z} - f_{h_V,V,z} : z \in \{0,1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n}),$$

is a Lipschitz transform of function sets $\mathcal{V}$, $\mathcal{V}^*$, $\mathcal{B}$, $\mathcal{G}$, $\mathcal{M}$, with bounded Lipschitz coefficients and with a constant envelope. Therefore, we have that

$$\log \sup_Q N(\varepsilon, \mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0,$$

Applying Lemma A.3 and Markov inequality, we have for some constant $K > e$

$$\begin{aligned}
&\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |II_V(z)| \leqslant \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)| \\
&= O_{\mathrm{P}}(1) \left(\sqrt{s\sigma_n^2 \log(p \vee K \vee \sigma_n^{-1})} + \frac{s}{\sqrt{n}} \log(p \vee K \vee \sigma_n^{-1}) \right) \\
&= O_{\mathrm{P}}(1) \left(\sqrt{s\delta_n^2 n^{-1/2} \log(p \vee n)} + \sqrt{s^2 n^{-1} \log^2(p \vee n)} \right) \\
&= O_{\mathrm{P}}(1) \left(\delta_n \delta_n^{1/4} + \delta_n^{1/2} \right) = O_{\mathrm{P}}(\delta_n^{1/2}),
\end{aligned}$$

for $\sigma_n = \sup_{f \in \mathcal{F}_n} \|f\|_{P,2}$; and we used some simple calculations, exploitng the boundedness conditions in Assumptions 2 and 3, to deduce that,

$$\sigma_n = \sup_{f \in \mathcal{F}_n} \|f\|_{P,2} \lesssim \sup_{h \in \mathcal{H}_{V,n}, V \in \mathcal{V}} \|h - h_V\|_{P,2} \lesssim \delta_n n^{-1/4}.$$

since $\sup_{h \in \mathcal{H}_{V,n}, V \in \mathcal{V}} \|h - h_V\|_{P,2} \lesssim \delta_n n^{-1/4}$ by definition of the set $\mathcal{H}_{V,n}$; and then we used that $s^2 \log^3(p \vee n)/n \leqslant \delta_n$ by Assumption 3.

(f) The claim of Step 1 follows by collecting steps (a)-(e).

Step 2 (Uniform Donskerness). Here we claim that Assumption 2 implies two assertions:

(a) The set of vector functions $(\psi_u^\rho)_{u \in \mathcal{U}}$, where $\psi_u^\rho := (\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\})_{V \in V_u}$, is P-Donsker uniformly in $\mathcal{P}$, namely that

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P},$$

where $Z_{n,P} := (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}}$ and $Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}$.

(b) Moreover, Z_P has bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{u \in \mathcal{U}} \|Z_P(u)\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u,\tilde{u}) \leqslant \varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| = 0.$$

To verify (a), we shall invoke Lemma A.2.

To demonstrate the claim, it will suffice to consider the set of $\mathbb{R}$-valued functions $\Psi = (\psi_{uk}, u \in \mathcal{U}, k \in 1, ..., d_\rho)$. Further, we notice that $\mathbb{G}_n \psi_{V,z}^\alpha = \mathbb{G}_n f$, for $f \in \mathcal{F}_z$,

$$\mathcal{F}_z = \left\{ \frac{1\{Z = z\}(V - g_V(z,X))}{m(z,X)} + g_V(z,X), V \in \mathcal{V} \right\},$$

and that $\mathbb{G}_n\psi_V^\gamma = \mathbb{G}_n f$, for $f = V \in \mathcal{V}$. Hence $\mathbb{G}_n(\psi_{uk}) = \mathbb{G}_n(f)$ for $f \in \mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{V}$. That Ψ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$ is then implied by $\mathcal{F}$ being Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$, so we demonstrate the latter claims below.

Observe that $\mathcal{F}_z$ is formed as a unform Lipschitz transform of function sets $\mathcal{B}, \mathcal{V}, \mathcal{V}^*, \mathcal{M}$ where validity of the Lipschitz property relies on Assumption 2 (to keep the denominator away from zero) and on boundedness conditions in Assumption 3. The latter function sets are uniformly bounded classes that have the uniform covering ϵ-entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant, and so this class, which is uniformly bounded under Assumption 2, has the uniform ϵ-entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant (e.g. van der Vaart and Wellner (1996)). Since $\mathcal{F}$ is uniformly bounded and is a finite union of function sets with the uniform entropies obeying the said properties, it also follows that it has this property, namely:

$$\log \sup_Q N(\varepsilon, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim C \log(\mathrm{e}/\epsilon) \vee 0.$$

Since $\int_0^\infty \sqrt{\log(\mathrm{e}/\epsilon) \vee 0} d\epsilon = \mathrm{e}\sqrt{\pi}/2 < \infty$ and $\mathcal{F}$ is uniformly bounded, application of Lemma A.2 then establishes that $\mathcal{F}$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$.[15]

To demonstrate claim (b), we need to translate pre-Gaussianity of $\mathcal{F}$ uniformly in $P \in \mathcal{P}$ into the continuity property stated above. One implication is simple:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{u\in\mathcal{U}} \|Z_P(u)\| \leqslant \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{f\in\mathcal{F}} \|\mathbb{G}_P(f)\| < \infty.$$

Consider a sequence of positive constants ε approaching zero, and note that

$$\mathrm{E}_P \sup_{d_\mathcal{U}(u,\tilde{u})\leqslant\varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| \lesssim \mathrm{E}_P \sup_{d_\mathcal{U}(u,\tilde{u})\leqslant\varepsilon} \max_{k\leqslant d_\rho} |\mathbb{G}_P(\psi_{uk} - \psi_{\bar{u}k})| \lesssim \mathrm{E}_P \sup_{d_\mathcal{U}(u,\tilde{u})\leqslant\varepsilon} |\mathbb{G}_P(f_u - f_{\bar{u}})|$$

where f_u and $f_{\bar{u}}$ must be of the form:

$$\frac{1\{Z = z\}(U_u - g_{U_u}(z,X))}{m(z,X)} + g_{U_u}(z,X), \frac{1\{Z = z\}(U_{\bar{u}} - g_{U_{\bar{u}}}(z,X))}{m(z,X)} + g_{U_{\bar{u}}}(z,X),$$

with $(U_u, U_{\bar{u}})$ equal to either $(Y_u, Y_{\bar{u}})$ or $(1_d(D)Y_u, 1_d(D)Y_{\bar{u}})$, for $d = 0$ or 1, and $z = 0$ or 1. Then

$$\sup_{P\in\mathcal{P}} \|f_u - f_{\bar{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}} \|Y_u - Y_{\bar{u}}\|_{P,2} \to 0,$$

as $d_\mathcal{U}(u,\bar{u}) \to 0$ by Assumption 2, which together with $\mathcal{F}$ being uniformly pre-Gaussian implies that:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_\mathcal{U}(u,\tilde{u})\leqslant\varepsilon} |\mathbb{G}_P(f_u - f_{\bar{u}})| \to 0,$$

which implies

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_\mathcal{U}(u,\tilde{u})\leqslant\varepsilon} \|Z_P(u) - Z_P(\tilde{u})\| \to 0.$$

[15]The set of functions $\mathcal{F}$ is image-admissible Suslin, since it is formed by a continuous transformation of $(Y_u, u \in \mathcal{U})$, which is image-admissible Suslin and a finite set of random variables (which are trivially image-admissible Suslin).

Lastly, $\sup_{P\in\mathcal{P}}\|f_u - f_{\bar{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}}\|Y_u - Y_{\bar{u}}\|_{P,2}$ follows from the fact that $\|f_u - f_{\bar{u}}\|_{P,2} \lesssim \|Y_u - Y_{\bar{u}}\|_{P,2}$ for each $P \in \mathcal{P}$, which follows from a sequence of inequalities holding for each $P \in \mathcal{P}$: (1)

$$\|f_u - f_{\bar{u}}\|_{P,2} \lesssim \|U_u - U_{\bar{u}}\|_{P,2} + \|g_{U_u}(z,X) - g_{U_{\bar{u}}}(z,X)\|_{P,2},$$

which we deduced using triangle inequality and the fact that $m(z,X)$ is bounded away from zero, (2) $\|U_u - U_{\bar{u}}\|_{P,2} \leqslant \|Y_u - Y_{\bar{u}}\|_{P,2}$, which we deduced using a Holder inequality, (3)

$$\|g_{U_u}(z,X) - g_{U_{\bar{u}}}(z,X)\|_{P,2} \leqslant \|U_u - U_{\bar{u}}\|_{P,2},$$

which we deduced by the definition of $g_V(z,X) = \mathrm{E}_P(V|X, Z = z)$ and the contraction property of conditional expectation recalled in Lemma A.6. ∎

B.2. **Proof of Theorem 4.2.** The proof will be similar to the previous proof, and as in that proof we only focus the presentation on the first strategy.

Step 0. (A Preamble). In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions only, but not on n once $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$, and not on $\mathrm{P} \in \mathcal{P}_n$. We consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $\mathrm{P} = \mathrm{P}_n$ throughout the proof, *suppressing* the index n. Since the argument is asymptotic, we can just assume that $n \geqslant n_0$ in what follows.

Let $\mathbb{P}_n$ denote the measure that puts mass n^{-1} on points (ξ_i, W_i) for $i = 1, ..., n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_n f = n^{-1}\sum_{i=1}^n f(\xi_i, W_i)$.

Recall the we define the bootstrap draw as:

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \widehat{\psi}^\rho_u(W_i)\right)_{u\in\mathcal{U}} = \mathbb{G}_n(\xi\widehat{\psi}^\rho).$$

Step 1.(Linearization) In this step we establish the first claim, namely that

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) = Z^*_{n,P} + o_P(1) \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \tag{68}$$

where $Z^*_{n,P} := (\mathbb{G}_n \xi\psi^\rho_u)_{u\in\mathcal{U}}$. The components $(\sqrt{n}(\widehat{\gamma}^*_{V_{uj}} - \widehat{\gamma}_{V_{uj}}))_{u\in\mathcal{U}}$ of $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$ trivially have the linear representation (with no error) for each $j \in \mathcal{J}$. We only need to establish the claim for the empirical process $(\sqrt{n}(\widehat{\alpha}^*_{V_{uj}}(z) - \widehat{\alpha}_{V_{uj}}(z)))_{u\in\mathcal{U}}$ for $z \in \{0,1\}$, which we do in the steps below.

(a) As in the previous proof, we have that with probability at least $1 - \Delta_n$,

$$\widehat{h}_V \in \mathcal{H}_{V,n} := \{h = (\bar{g}(0,\cdot), \bar{g}(1,\cdot), \bar{m}(0,\cdot), \bar{m}(1,\cdot),) \in \mathcal{G}_V(0) \times \mathcal{G}_V(1) \times \mathcal{M}(0) \times \mathcal{M}(1)\}.$$

(b) We have that

$$\sqrt{n}(\widehat{\alpha}_V(z) - \alpha_V(z)) = \underbrace{\mathbb{G}_n[\xi f_{h_V,V,z}]}_{I^*_V(z)} + \underbrace{(\mathbb{G}_n[\xi f_{h,V,z}] - \mathbb{G}_n[\xi f_{h_V,V,z}])}_{II^*_V(z)} + \underbrace{\sqrt{n}(\mathrm{E}[\xi f_{h,V,z} - \xi f_{h_V,h,z}])}_{III^*_V(z)},$$

with h evaluated at $h = \widehat{h}_V$.

(c) Note that $III_V^*(z) = III_V(z)$ since ξ is independent of W, so by the previous proof since $\widehat{h}_V \in \mathcal{H}_{V,n}$ for all $V \in \mathcal{V}$ with probability $1 - \Delta_n$, we have that once $n \geqslant n_0$,

$$\mathrm{P}_P\Big(|III_V^*(z)| \lesssim \delta_n^2, \forall z \in \{0,1\}, \forall V \in \mathcal{V}\Big) \geqslant 1 - \Delta_n.$$

(d). Furthermore, we have that

$$\sup_{V\in\mathcal{V}} \max_{z\in\{0,1\}} |II_V(z)| \leqslant \sup_{h\in\mathcal{H}_{V,n}, z\in\{0,1\}, V\in\mathcal{V}} |\mathbb{G}_n[\xi f_{h,V,z}] - \mathbb{G}_n[\xi f_{h_V,V,z}]|.$$

By the previous proof the class of functions, $\mathcal{F}_n = (f_{h,V,z} - f_{h_V,V,z} : z \in \{0,1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n})$, obeys $\log \sup_Q N(\varepsilon, \mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0$. By Lemma A.5, multiplication of this class with ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\varepsilon, \xi\mathcal{F}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(1/\varepsilon)) \vee 0,$$

since the envelope for $\xi\mathcal{F}_n$ this class is $|\xi|$ times a constant, and $\mathrm{E}[\xi^2] = 1$. We also have that, by standard calculations,

$$(\mathrm{E}[\max_{i\leqslant n} |\xi|^2])^{1/2} \lesssim \log n.$$

Applying Lemma A.3 and Markov inequality, we have for some constant $K > e$

$$\begin{aligned}
&\sup_{V\in\mathcal{V}} \max_{z\in\{0,1\}} |II_V(z)| \leqslant \sup_{f\in\mathcal{F}_n} |\mathbb{G}_n(f)| \\
&= O_{\mathrm{P}}(1)\left(\sqrt{s\sigma_n^2 \log(p \vee K \vee \sigma_n^{-1})} + \frac{s \log n}{\sqrt{n}} \log(p \vee K \vee \sigma_n^{-1})\right) \\
&= O_{\mathrm{P}}(1)\left(\sqrt{s\delta_n^2 n^{-1/2} \log(p \vee n)} + \sqrt{s^2 n^{-1} \log^3(p \vee n)}\right) \\
&= O_{\mathrm{P}}(1)\left(\delta_n \delta_n^{1/4} + \delta_n^{1/2}\right) = O_{\mathrm{P}}(\delta_n^{1/2}),
\end{aligned}$$

for $\sigma_n = \sup_{f\in\xi\mathcal{F}_n} \|f\|_{P,2} = \sup_{f\in\mathcal{F}_n} \|f\|_{P,2}$; where the details of calculations are the same as in the previous proof.

(e) The claim of Step 1 follows by collecting steps (a)-(d).

Step 2 (Unconditional Uniform Donskerness). Here we claim that Assumption 2 implies: The set of vector functions $(\xi\psi_u^\rho)_{u\in\mathcal{U}}$, where $\psi_u^\rho := (\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\})_{V\in V_u}$, is P-Donsker uniformly in $\mathcal{P}$, namely that

$$Z_{n,P}^* \rightsquigarrow Z_P^* \quad \text{in } \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P},$$

where $Z_{n,P}^* := (\mathbb{G}_n \xi\psi_u^\rho)_{u\in\mathcal{U}}$ and $Z_P^* := (\mathbb{G}_P \xi\psi_u^\rho)_{u\in\mathcal{U}}$ is equal in distribution to $Z_P := (\mathbb{G}_P \xi\psi_u^\rho)_{u\in\mathcal{U}}$, in particular, Z_P^* and Z_P share the identical covariance function.

To verify (a), we shall invoke Lemma A.2. To demonstrate the claim, it will suffice to consider the set of $\mathbb{R}$-valued functions $\xi\Psi$, where $\Psi = (\psi_{uk}, u \in \mathcal{U}, k \in 1, ..., d_\rho)$. As in the previous proof, we notice that $\mathbb{G}_n(\xi\psi_{uk}) = \mathbb{G}_n(\xi f)$ for $f \in \mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{V}$. That $\xi\Psi$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$ is then implied by $\xi\mathcal{F}$ being Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$, so we demonstrate the latter claim. From the previous proof, $\log \sup_Q N(\varepsilon, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$. By Lemma A.5 multiplication by ξ does not change the entropy bound, modulo a multiplicative

constant, hence $\log \sup_Q N(\varepsilon, \xi\mathcal{F}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$. This establishes that $\xi\mathcal{F}$ is Donsker and pre-Gaussian uniformly in $P \in \mathcal{P}$. Since multiplication by ξ to create $\mathbb{G}_P(\xi f)$ does not change the covariance function of $\mathbb{G}_P(f)$, the P-Gaussian processes indexed by $\xi\mathcal{F}$ and by $\mathcal{F}$ are equal in distribution. The claim that $Z^*_{n,P} \rightsquigarrow Z_P$ then follows as in the previous proof.

Step 3 (Uniform Donskerness Conditional on Data). The previous argument implies unconditional convergence in distribution under any sequence $P = P_n \in \mathcal{P}_n$. Using the same argument as in the first part of the proof of Theorem 2.9.6 in van der Vaart and Wellner (1996), we can claim that the conditional convergence takes place under any sequence $P = P_n \in \mathcal{P}_n$, using the unconditional convergence to establish stochastic equicontinuity for the conditional convergence. Moreover, linearization error in Step 1 converges to zero in unconditional probability. It is known that this is stronger than the conditional convergence. The final claim follows by combining the steps. ∎

B.3. **Proof of Theorem 4.3.** We have that under the sequence P_n

$$Z_{P_n,n} \rightsquigarrow Z_{P_n},$$

which means that $\lim_{n\to\infty} \sup_{h\in BL_1} |\mathrm{E}^*_{P_n} h(Z_{P_n,n}) - \mathrm{E}_{P_n} h(Z_{P_n})| = 0$. By the uniform in $P \in \mathcal{P}$ pre-Gaussianity of Z_{P_n} and compactness of $\mathbb{D}_\rho$ we can split $\mathbb{Z}$ into a collection of subsequences $\{\mathbb{Z}'\}$, along each of which

$$Z_{P_n} \rightsquigarrow Z', \quad \rho_{P_n} \to \rho',$$

meaning that $\lim_{n\in\mathbb{Z}'} \sup_{h\in BL_1} |\mathrm{E}_{P_n} h(Z_{P_n}) - \mathrm{E}_{P_n} h(Z')| = 0$, where Z' is a tight Gaussian process, which depends on a subsequence $\mathbb{Z}'$ with paths that are continuous on $\mathcal{U}$, with covariance function equal to the limit of the covariance function Z_{P_n} along the subsequence, which may depend on $\mathbb{Z}'$, and ρ' is some value that also depends on the subsequence. We can conclude by the triangle inequality that along that same subsequence,

$$Z_{P_n,n} \rightsquigarrow Z'.$$

Application of the functional delta method for subsequences, Lemma A.7, yields

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\rho'}(Z')$$

and, furthermore, by Assumption 5 and the extended continuous mapping theorem,

$$\phi'_{\rho_{\mathbb{P}_n}}(Z_{P_n}) \rightsquigarrow \phi'_{\rho'}(Z').$$

Since the argument above works for all subsequences as defined above, we conclude that

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P), \quad \text{in } \ell^\infty(\mathcal{W})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n.$$

The conclusion for bootstrap follows similarly, except now we apply Lemma A.8. ∎

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Andrews, D. W. (1994): "Empirical process methods in econometrics," *Handbook of Econometrics*, 4, 2247–2294.

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

Bach, F. (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429, Arxiv, 2010.

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

——— (2011b): "ℓ_1-penalized quantile regression in high-dimensional sparse models," *Ann. Statist.*, 39(1), 82–130.

——— (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

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

——— (2011): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *ArXiv*, forthcoming, The Review of Economic Studies.

——— (2013): "Inference for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III, 245–295.

Belloni, A., V. Chernozhukov, and K. Kato (2013): "Uniform Post Selection Inference for LAD Regression Models," *arXiv preprint arXiv:1304.0282*.

Belloni, A., V. Chernozhukov, and L. Wang (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806, Arxiv, 2010.

Belloni, A., V. Chernozhukov, and Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv preprint arXiv:1304.3969*.

Benjamin, D. J. (2003): "Does 401(k) eligibility increase saving? Evidence from propensity score subclassification," *Journal of Public Economics*, 87, 1259–1290.

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

Candès, 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. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chamberlain, G., and G. W. Imbens (2003): "Nonparametric applications of Bayesian inference," *Journal of Business & Economic Statistics*, 21(1), 12–18.

Chernozhukov, V., and C. Hansen (2004): "The impace of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis," *Review of Economics and Statistics*, 86(3), 735–751.

Dudley, R. (2000): *Uniform Cental Limit Theorems*. Cambridge Studies in advanced mathematics.

Engen, E. M., and W. G. Gale (2000): "The Effects of 401(k) Plans on Household Wealth: Differences Across Earnings Groups," Working Paper 8032, National Bureau of Economic Research.

Engen, Eric M., W. G. G., and J. K. Scholz (1996): "The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives*, 10, 113–138.

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. (2013): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," .

Frank, I. E., and J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35(2), 109–135.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, pp. 315–331.

Heckman, J., and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proc. Natl. Acad. Sci. USA*, 96(8), 4730–4734 (electronic).

Hong, H., and D. Nekipelov (2010): "Semiparametric efficiency in nonlinear LATE models," *Quantitative Economics*, 1, 279–304.

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

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

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kato, K. (2011): "Group Lasso for high dimensional sparse quantile regression models," Preprint, ArXiv.

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

——— (2005): *Quantile regression*, no. 38. Cambridge university press.

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

——— (2008b): "Recent developments in model selection and related areas," *Econometric Theory*, 24(2), 319–322.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, pp. 255–285.

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.

Neyman, J. (1979): "$C(\alpha)$ tests and their use," *Sankhya*, 41, 1–21.

Poterba, J. M., S. F. Venti, and D. A. Wise (1994): "401(k) Plans and Tax-Deferred savings," in *Studies in the Economics of Aging*, ed. by D. A. Wise. Chicago, IL: University of Chicago Press.

——— (1995): "Do 401(k) Contributions Crowd Out Other Personal Saving?," *Journal of Public Economics*, 58, 1–32.

——— (1996): "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence," Working Paper 5599, National Bureau of Economic Research.

——— (2001): "The Transition to Personal Accounts and Increasing Retirement Wealth: Macro and Micro Evidence," Working Paper 8610, National Bureau of Economic Research.

Pötscher, B. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhya*, 71-A, 1–18.

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

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

Tsybakov, A. B. (2009): *Introduction to nonparametric estimation*. Springer.

van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

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

Vytlacil, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

Wasserman, L. (2006): *All of nonparametric statistics*, vol. 4. Springer New York.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, second edn.

Zou, H. (2006): "The Adaptive Lasso And Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Table 1: Estimates and standard errors of average effects

Specification					
Series approximation	Dimension	Selection	Selected (Main effects)	LATE	LATE-T
Indicators	20	N		11833	16120
				(1638)	(2224)
				{1764}	{2393}
Indicators	20	Y	13	12382	16419
				(1684)	(2205)
				{1694}	{2224}
Indicators plus interactions	167	N		11856	16216
				(1632)	(2224)
				{1625}	{2211}
Indicators plus interactions	167	Y	31 (9)	12981	16957
				(1702)	(2183)
				{1663}	{2128}
B-splines	27	N		11559	15591
				(1571)	(2135)
				{1741}	{2359}
B-splines	27	Y	14	11925	15557
				(1594)	(2188)
				{1583}	{2175}
B-splines plus interactions	300	N		35792	73053
				(14394)	(37609)
				{13873}	{35877}
B-splines plus interactions	300	Y	37 (8)	12134	15547
				(1580)	(2209)
				{1549}	{2215}

Notes: The sample is drawn from the 1991 SIPP and consists of 9,915 observations. All the specifications control for age, income, family size, education, marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status. Indicators specification uses a linear term for family size, 5 categories for age, 4 categories for education, and 7 categories for income. B-splines specification uses b-splines with 3, 4, 6, and 8 knots for family size, education, age, and income, respectively. Marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status are included as indicators in all the specifications. Analytical standard errors are given in parentheses. Weighted bootstrap standard errors based on 500 repetitions with standard exponential weights are given in braces.

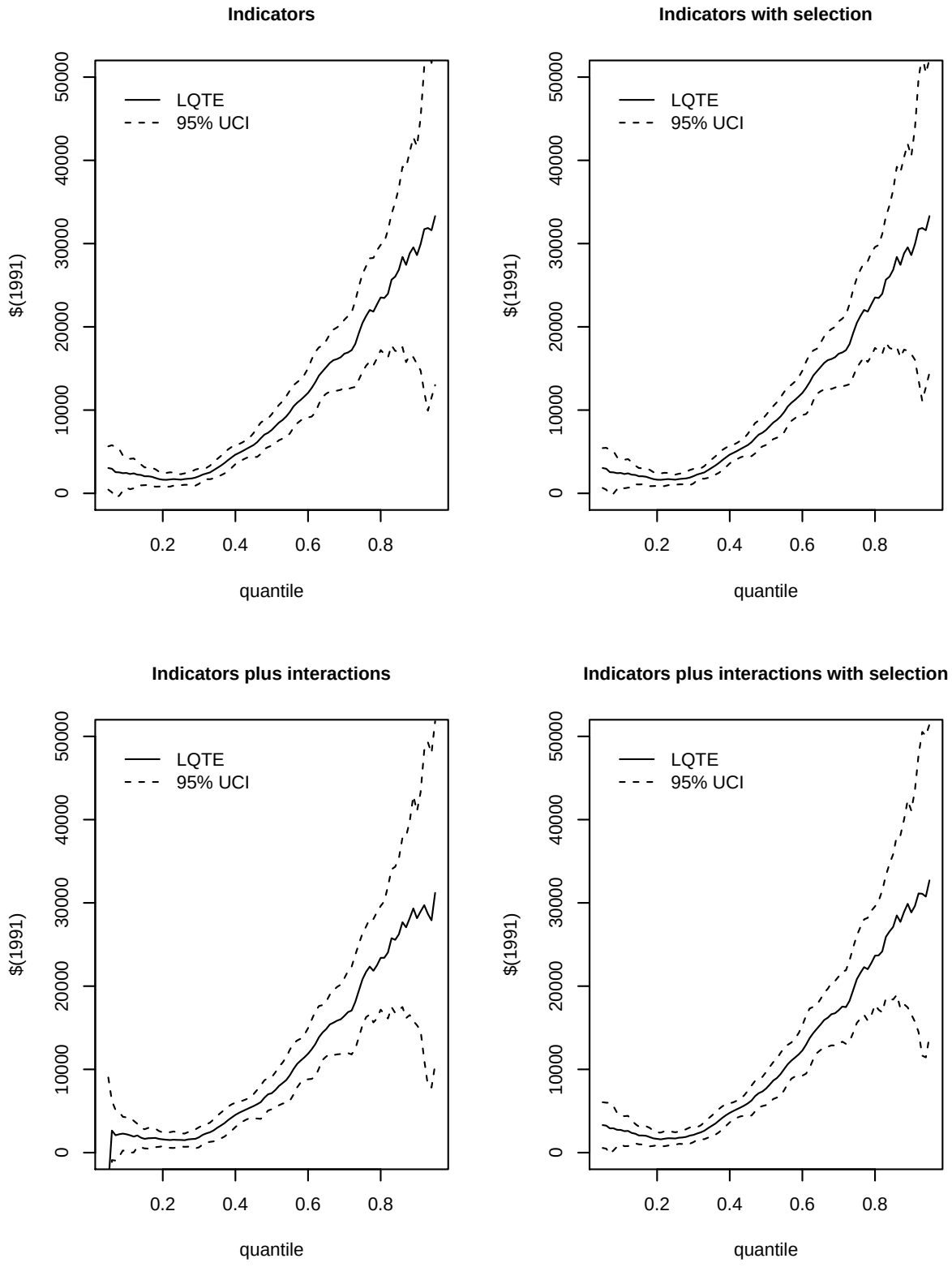

FIGURE 1. LQTE estimates based on indicator specification.

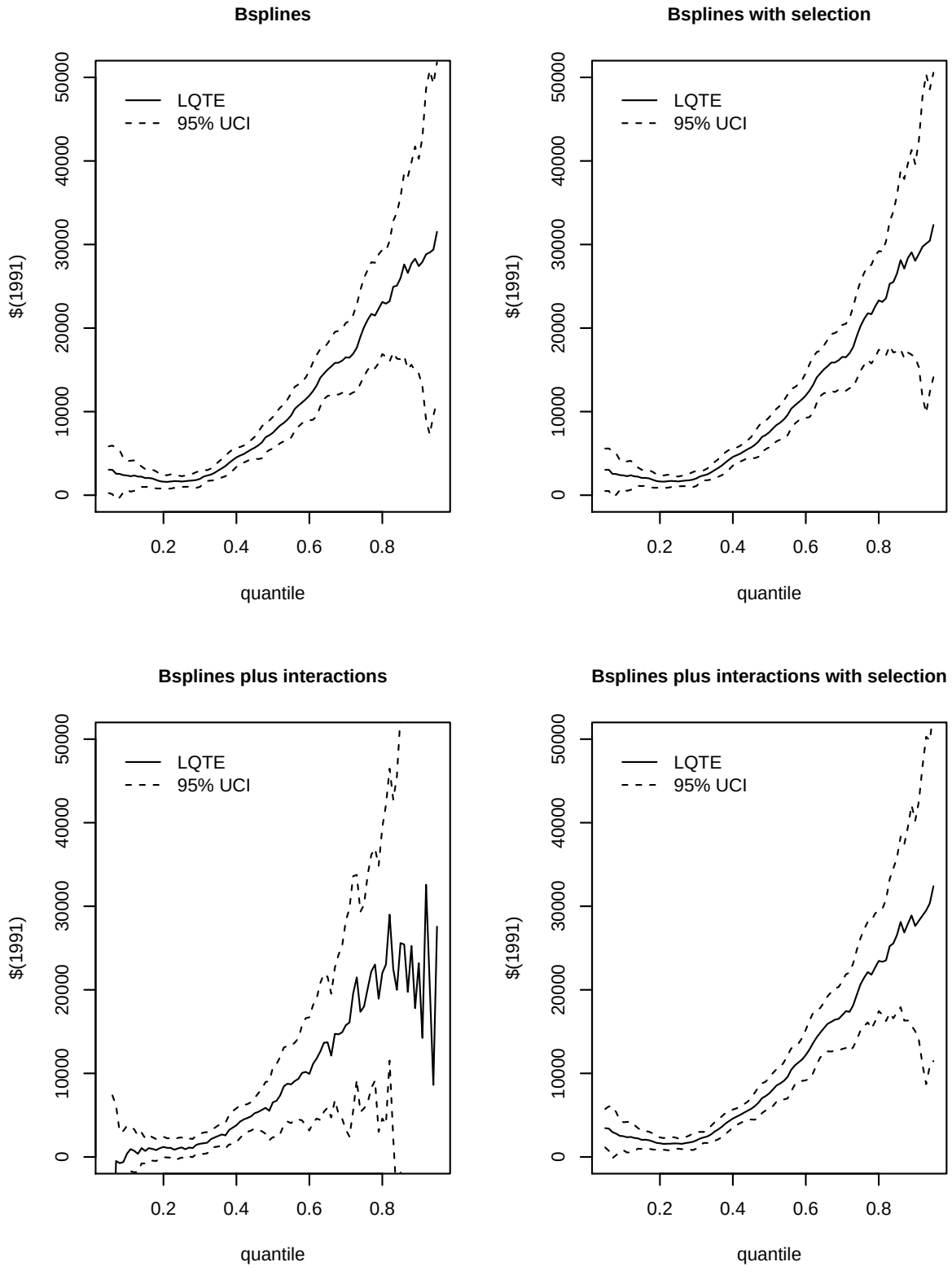

FIGURE 2. LQTE estimates based on b-spline specification.

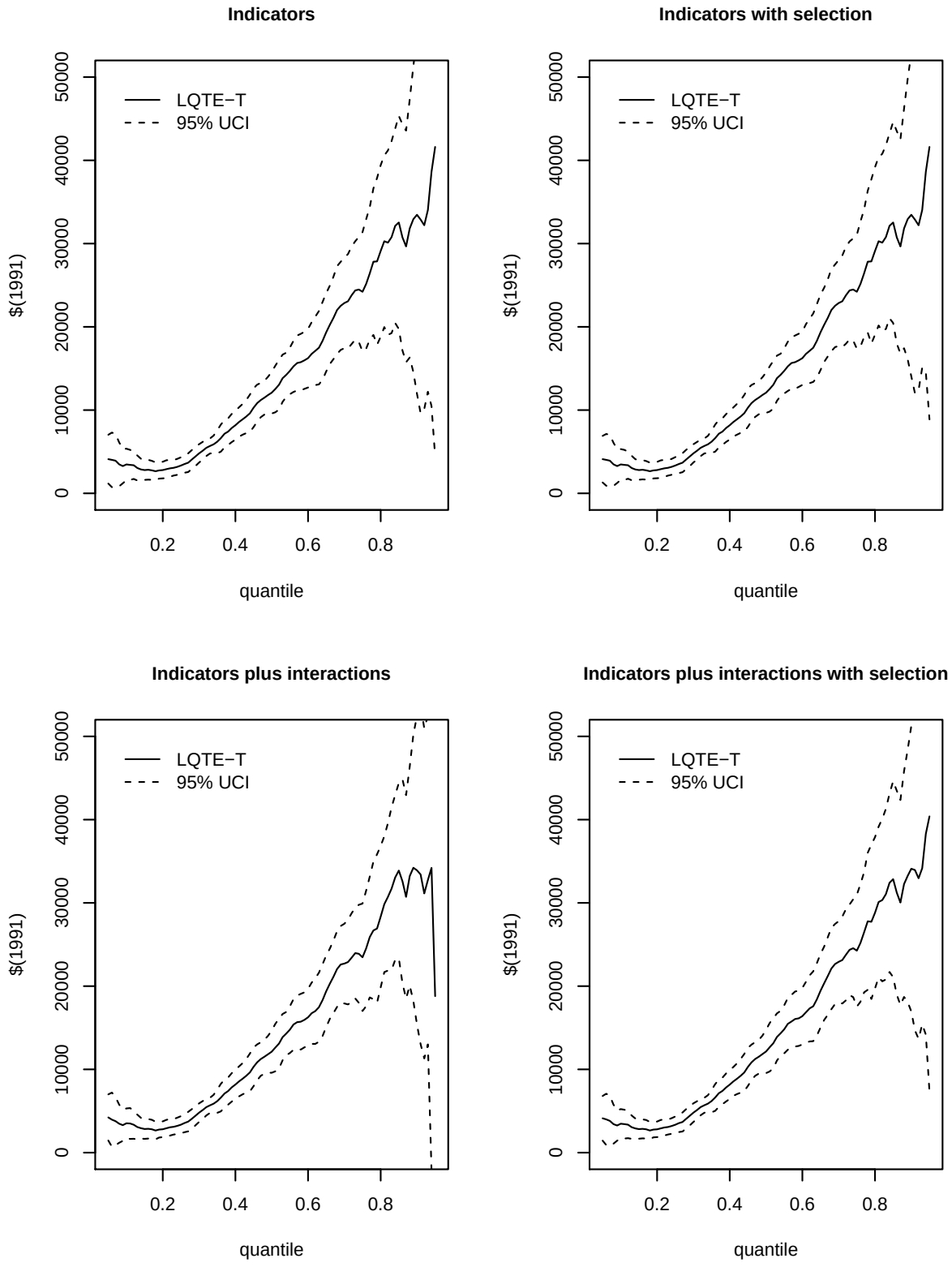

FIGURE 3. LQTE-T estimates based on indicator specification.

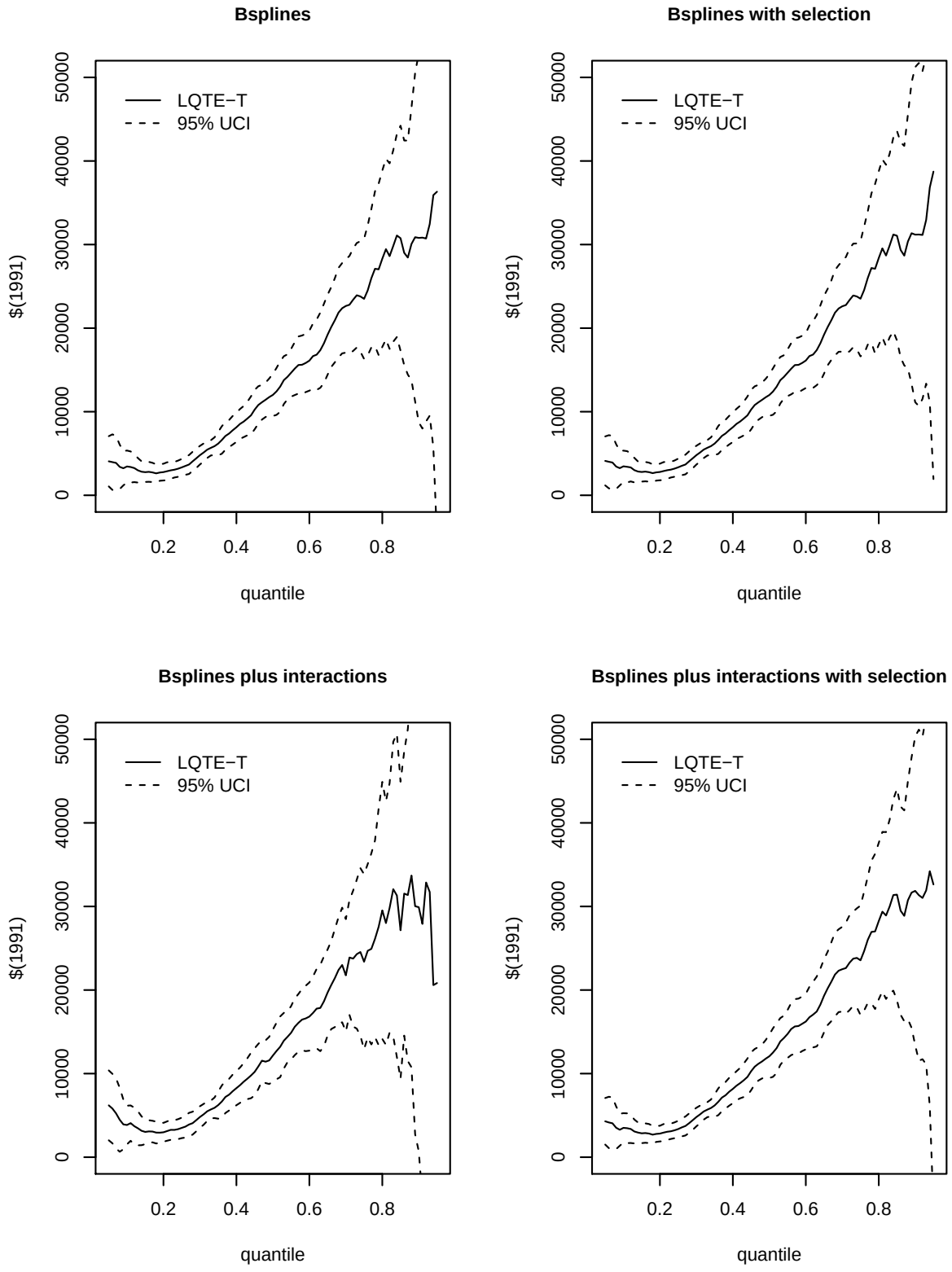

FIGURE 4. LQTE-T estimates based on b-spline specification.