Econometrica, Vol. 85, No. 1 (January, 2017), 233–298

PROGRAM EVALUATION AND CAUSAL INFERENCE WITH HIGH-DIMENSIONAL DATA

BY A. BELLONI, V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN[1]

In this paper, we provide efficient estimators and honest confidence bands for a variety of treatment effects including local average (LATE) and local quantile treatment effects (LQTE) in data-rich environments. We can handle *very many* control variables, *endogenous* receipt of treatment, *heterogeneous* treatment effects, and *function-valued* outcomes. Our framework covers the special case of exogenous receipt of treatment, either conditional on controls or unconditionally as in randomized control trials. In the latter case, our approach produces efficient estimators and honest bands for (functional) average treatment effects (ATE) and quantile treatment effects (QTE). To make informative inference possible, we assume that key reduced-form predictive relationships are approximately sparse. This assumption allows the use of regularization and selection methods to estimate those relations, and we provide methods for post-regularization and post-selection inference that are uniformly valid (honest) across a wide range of models. We show that a key ingredient enabling honest inference is the use of orthogonal or doubly robust moment conditions in estimating certain reduced-form functional parameters. We illustrate the use of the proposed methods with an application to estimating the effect of 401(k) eligibility and participation on accumulated assets.

The results on program evaluation are obtained as a consequence of more general results on honest inference in a general moment-condition framework, which arises from structural equation models in econometrics. Here, too, the crucial ingredient is the use of orthogonal moment conditions, which can be constructed from the initial moment conditions. We provide results on honest inference for (function-valued) parameters within this general framework where *any high-quality*, *machine learning* methods (e.g., boosted trees, deep neural networks, random forest, and their aggregated and hybrid versions) can be used to learn the nonparametric/high-dimensional components of the model. These include a number of supporting auxiliary results that are of major independent interest: namely, we (1) prove uniform validity of a multiplier bootstrap, (2) offer a uniformly valid functional delta method, and (3) provide results for sparsity-based estimation of regression functions for function-valued outcomes.

KEYWORDS: Machine learning, causality, Neyman orthogonality, heterogenous treatment effects, endogeneity, local average and quantile treatment effects, instruments, local effects of treatment on the treated, propensity score, Lasso, inference after model selection, moment-condition models, moment-condition models with a continuum of target parameters, Lasso and Post-Lasso with functional response data, randomized control trials.

1. INTRODUCTION

THE GOAL OF MANY EMPIRICAL ANALYSES is to understand the causal effect of a treatment, such as participation in a training program or a government policy, on economic

[1]We gratefully acknowledge research support from the NSF. We are very grateful to the co-editor, three anonymous referees, Alberto Abadie, Stephane Bonhomme, Matias Cattaneo, Jinyong Hahn, Michael Jansson, Toru Kitagawa, Roger Koenker, Simon Lee, Yuan Liao, Oliver Linton, Blaise Melly, Whitney Newey, Adam Rosen, and seminar participants at the Bristol Econometric Study Group, Cambridge, CEMFI, Cornell–Princeton Conference on "Inference on Non-Standard Problems," Winter 2014 ES meeting, Semi-Plenary Lecture at of ES Summer Meeting 2014, ES World Congress 2015, 2013 Summer NBER Institute, University of Montreal, UIUC, UCL, and University of Warwick for helpful comments. We are especially grateful to Andres Santos for many useful comments. Matlab code for the empirical illustration is available on the *Econometrica* website. R code for implementing the treatment effects estimators is available in the R package "hdm" from Spindler, Chernozhukov, and Hansen (2016).

 DOI: 10.3982/ECTA12723

and other outcomes. Such analyses are often complicated by the fact that treatments or policies are rarely 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 approaches include instrumental variable (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 (or simply controls). Another common approach is to assume that the treatment variable itself may be taken as exogenous after conditioning on the right set of controls which leads to regression or matching based methods, among others, for estimating treatment effects.[2]

A practical problem empirical researchers 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 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 controls 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 controls from a potentially large set including raw variables available in the data as well as interactions and other transformations of these variables.

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 potential controls, $f(X)$. Allowing many potential controls expressly covers both the case where there are simply many controls (where $f(X) = X$) and the case where there are many technical controls $f(X)$ generated as transformations such as powers, b-splines, or interactions of raw controls,[3] X, along with combinations of the two cases. The notation $f(X)$ naturally accommodates these cases, and we call $f(X)$ the *controls* regardless of the case. 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 *endogenous* case where identification is obtained through the use of an instrumental variable, but all results carry through to the *exogenous* case where the treatment is taken as exogenous unconditionally or after conditioning on sufficient controls by simply replacing the instrument with the treatment variable in the estimation and inference methods and in the formal results. In the latter case, LATE reduces to the average treatment effect (ATE) and LQTE to the quantile treatment effect (QTE).

The methodology for estimating treatment effects we consider allows for cases where the number of potential controls, $p := \dim f(X)$, is much larger than the sample size,

[2]There is a large literature about estimation of treatment effects. See, for example, the textbook treatments in Angrist and Pischke (2008), Wooldridge (2010), and Imbens and Rubin (2015).

[3]See, for example, Koenker (1988), Newey (1997), Wasserman (2006), Chen (2007), and Tsybakov (2009).

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 the conditional expectations $\mathrm{E}_P[D|X]$, $\mathrm{E}_P[Z|X]$, and $\mathrm{E}_P[Y|X]$ 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 number $s \ll n$ of the variables in $f(X)$ whose identities are a priori unknown to the researcher. This assumption allows us to use methods for estimating models in high-dimensional sparse settings 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 nontrivial. 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 can be restrictive and seems unlikely to be satisfied in many economic applications. For example, a typical condition that allows perfect model selection is the "beta-min" condition, which requires that all but a small number of coefficients are exactly zero and that the nonzero coefficients are all large enough that they can be distinguished from zero with probability very near 1 in finite samples. Such a condition rules out the possibility that there may be some variables which have moderate, but nonzero, partial effects. Ignoring such variables may result in large omitted variables bias that has a substantive impact on estimation and inference regarding individual model parameters; see Leeb and Pötscher (2008a, 2008b), Pötscher (2009), and Belloni, Chernozhukov, and Hansen (2013, 2014).

The *first main contribution* of this 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 Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012), who were the first to demonstrate, in a highly specialized context, that valid inference can proceed following model selection allowing for model selection mistakes under two conditions. We formulate and extend these two conditions to a rather general moment-condition framework (e.g., Hansen (1982) and Hansen and Singleton (1982)) as follows. First, estimation should be based upon "orthogonal" moment conditions that are first-order insensitive to changes in the values of nuisance parameters that will be estimated using high-dimensional methods. Specifically, if the target parameter value α_0 is identified via the moment condition

$$\mathrm{E}_P \psi(W, \alpha_0, h_0) = 0, \tag{1.1}$$

where h_0 is a function-valued nuisance parameter estimated via a model selection or regularization method, one needs to use a moment function, ψ, such that the corresponding moment condition is orthogonal with respect to perturbations of h around h_0. More formally, the moment condition should satisfy the *Neyman orthogonality condition*:

$$\partial_h \big[\mathrm{E}_P \psi(W, \alpha_0, h)\big]_{h=h_0} = 0, \tag{1.2}$$

where ∂_h is a functional derivative operator with respect to h restricted to directions of possible deviations of estimators of h_0 from h_0. Second, one needs to ensure that the model selection mistakes occurring in the estimation of nuisance parameters are uniformly "moderately" small with respect to the underlying model. Specifically, we will require that the nuisance parameter h_0 is estimated at the rate $o(n^{-1/4})$, which ensures small bias, and that the estimator takes values in a space whose entropy does not grow too fast, which ensures no overfitting. In this paper, we establish that building estimators based upon moment conditions with the orthogonality condition (1.2) holding ensures that crude estimation of h_0 via post-selection or other regularization methods has an asymptotically negligible effect on the estimation of α_0 in general frameworks. It then follows that we can form a regular, root-n consistent estimator of α_0, uniformly with respect to the underlying model.

In the endogenous treatment effects setting, we build moment conditions satisfying (1.2) from the efficient influence functions for certain reduced-form parameters, building upon Hahn (1998). We illustrate how orthogonal moment conditions coupled with methods developed for forecasting in high-dimensional approximately sparse models can be used to estimate and obtain valid inferential statements about a wide variety of structural/treatment effects. We formally demonstrate the uniform validity of the resulting inference within a broad class of approximately sparse models including models where perfect model selection is theoretically impossible. An important feature of our main theoretical results is that they cover the use of variable selection for *functional response data* using ℓ_1-penalized methods. Functional response data arise, for example, when one is interested in the LQTE at not just a single quantile but over a range of quantile indices. Considering this case then necessitates looking at the functional dependent variable $u \longmapsto 1(Y \leq u)$, where u denotes various levels that Y can cross. Treating such functional response data allows us to provide a unified inference procedure for interesting quantities such as the (local) distributional and quantile effects of the treatment, including simpler important parameters such as LQTE at a given quantile as a special case.

The *second main contribution* of this paper is providing a general set of results for uniformly valid estimation and inference methods in moment-condition problems, arising in structural analysis in econometrics and other data sciences. These results are useful not only for establishing the properties of treatment effects estimators developed here, but they are also useful for attacking a wide range of problems in structural econometrics. For example, Chernozhukov, Hansen, and Spindler (2015a) provided estimates of parameters characterizing a simple structural demand model based loosely on the analysis in Berry, Levinsohn, and Pakes (1995) using the framework developed here; see also Chernozhukov, Hansen, and Spindler (2015b). A key element to our establishing uniform validity of post-regularization inference is again the use of Neyman orthogonal moment conditions. In the general framework we consider, we may have (a continuum of) target parameters identified via (a continuum of) moment conditions that involve (a continuum of) nuisance functions that will be estimated via Lasso, Post-Lasso, or some other high-quality machine learning method. Our general theory expressly allows for a wide variety of traditional and machine learning methods, including those that do not rely on approximate sparsity, as long as the methods:

(1) have good approximation ability and
(2) do not overfit.

By "not overfitting" we mean that the entropy of the function classes containing the realizations of the estimator of the nuisance function/parameter does not increase too rapidly with the sample size. This second condition can only be verified analytically, but can be avoided by the use of various data splitting methods. For example, we can set aside a vanishing fraction of the data to estimate the nuisance parameter, as in Bickel (1982), or employ cross-fitting, as in Belloni, Chernozhukov, and Hansen (2010), Belloni et al. (2012) and Chernozhukov, Chetverikov, Demirer, Duflo, and Hansen (2016). Either scheme ensures that there is no asymptotic efficiency loss from data-splitting. We refer the reader to Chernozhukov et al. (2016) for a detailed discussion and analysis of cross-fitting in connection to inference on ATE and other causal parameters using machine learning methods for high-dimensional data.[4]

These results contain the results on treatment effects relevant for program evaluation, particularly the results for distributional and quantile effects, as a leading special case. These results are also immediately useful in other contexts such as nonseparable quantile models as in Chernozhukov and Hansen (2005, 2006), Chesher (2003), and Imbens and Newey (2009); semiparametric and partially identified models as in Escanciano and Zhu (2013); and many others. In our results, we first establish a functional central limit theorem for the continuum of target parameters and show that this functional central limit theorem holds uniformly in a wide range of data-generating processes P with approximately sparse continua of nuisance functions. Second, we establish a functional central limit theorem for the multiplier bootstrap that resamples the first-order approximations to the standardized estimators and demonstrate its uniform-in-P validity. These uniformity results build upon and complement those given in Romano and Shaikh (2012) for the empirical bootstrap. Third, we establish a functional delta method for smooth functionals of the continuum of target parameters and a functional delta method for the multiplier bootstrap of these smooth functionals, both of which hold uniformly in P, using an appropriately strengthened notion of Hadamard differentiability. All of these results are new and are of independent interest outside of the treatment effects focus of this paper.

We illustrate the use of our methods by estimating the effect of 401(k) eligibility and 401(k) participation on measures of accumulated assets as in Chernozhukov and Hansen (2004).[5] Similarly to Chernozhukov and Hansen (2004), we provide estimates of ATE and QTE of 401(k) eligibility and of LATE and LQTE of 401(k) participation. We differ from this previous work by using the high-dimensional methods developed in this paper to allow ourselves to consider a broader set of controls than has previously been considered. We find that 401(k) participation has a moderate 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. It is interesting that these results are similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of controls.

[4]Cross-fitting proceeds as follows: (1) split the sample into two equal parts, the auxiliary and main parts; (2) use the auxiliary part to estimate the nuisance parameter and the main part to estimate the target parameter, obtaining one estimator of the target parameter; (3) by reversing the roles of the main and auxiliary parts, obtain another estimator of the target parameter; and (4) average the two estimators of the target parameter to obtain the final estimator. The theorems established in Section 5 yield the properties of the final estimator; see Chernozhukov et al. (2016).

[5]See also Poterba, Venti, and Wise (1994, 1995, 1996, 2001), Abadie (2003), Benjamin (2003), and Ogburn, Rotnitzky, and Robins (2015), among others.

Links to the Literature

The Neyman orthogonality condition embodied in (1.2) 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. See also Newey (1990, 1994), Andrews (1994b), Robins and Rotnitzky (1995), and Linton (1996) for the use of this condition in semiparametric problems.

To the best of our knowledge, Belloni, Chernozhukov, and Hansen (2010) and Belloni et al. (2012) were the first to use the orthogonality (1.2) to expressly address the question of the uniform post-selection inference without imposing "beta-min" conditions, either in high-dimensional settings with $p \gg n$ or in low-dimensional settings with $p \ll n$. They applied it to the specific problem of the linear instrumental variables model with many instruments where the nuisance function h_0 is the optimal instrument estimated by Lasso or Post-Lasso methods and α_0 is the coefficient of the endogenous regressor. Belloni, Chernozhukov, and Hansen (2013, 2014) also exploited this approach to develop a double-selection method that yields valid post-selection inference on the parameters of the linear part of a partially linear model and on average treatment effects when the treatment is binary and *exogenous* conditional on controls in both the $p \gg n$ and the $p \ll n$ setting.[6] Subsequently, Farrell (2015) extended the results of Belloni, Chernozhukov, and Hansen (2013, 2014) to estimation of ATE when the treatment is multivalued and exogenous conditional on controls using group penalization for selection. Note that this previous work on treatment effects covers only the exogenous case and does not allow for functional responses which are necessary, for example, for working with distributional or quantile treatment effects.

Our work also contributes to the line of research on obtaining $\sqrt{n}$-consistent and asymptotically normal estimates for low-dimensional components within traditional semiparametric frameworks as in the important work by Bickel (1982), Robinson (1988), Newey (1990, 1994), van der Vaart (1991), Andrews (1994b), Ai and Chen (2003, 2012), and Chen, Linton, and Keilegom (2003). The major difference is that we allow for the use of modern high-dimensional methods, a.k.a. machine learning methods, for modeling and fitting the nonparametric (or high-dimensional) components of the model. In contrast to the former literature, we expressly allow for data-driven choice of the approximating model for the high-dimensional component, which addresses a crucial problem that arises in empirical work. Moreover, recent methods based on ℓ_1-penalization, upon which we focus in this paper, allow for much more flexible modeling of the nonparametric/high-dimensional parts of the model.[7] Our general theory in Section 5 also allows, in principle, for a wide variety of both traditional and machine learning methods.

The paper also generates a number of new results on sparse estimation with functional response data. These results are of independent interest in themselves, and they build upon the work of Belloni and Chernozhukov (2011) who provided 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

[6]Note that these results as well as results of this paper on the uniform post-selection inference in moment-condition problems are new for either $p \ll n$ or $p \gg n$ settings. The results also apply to arbitrary model selection devices, such as the Dantzig selector, Square-Root-Lasso, or Adaptive Lasso, that are able to select good sparse approximating models; and "moderate" model selection errors are explicitly allowed in the paper.

[7]See, for instance, Belloni et al. (2012) and Belloni, Chernozhukov, and Wang (2014) for a formalization of this claim in terms of rearranged Sobolev spaces where it is shown that traditional methods can fail to be consistent while ℓ_1-penalized methods remain consistent and have good rates of convergence.

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), van de Geer (2008), Huang, Horowitz, and Ma (2008), Bickel, Ritov, and Tsybakov (2009), Meinshausen and Yu (2009), Bach (2010), Huang, Horowitz, and Wei (2010), Belloni and Chernozhukov (2013, 2011), Kato (2011), Belloni et al. (2012), Belloni, Chernozhukov, and Kato (2013), Belloni, Chernozhukov, and Wei (2013), Caner and Zhang (2014), and the references therein.

Plan of the Paper

Section 2 introduces the structural parameters for policy evaluation and relates these parameters to reduced-form functions. Section 3 describes a three-step procedure to estimate and make inference on the structural parameters and functionals of these parameters, and Section 4 provides asymptotic theory in the treatment effects setting. Section 5 generalizes the setting and results to moment-condition problems with a continuum of structural parameters and a continuum of reduced-form functions. Section 6 derives general asymptotic theory for the Lasso and Post-Lasso estimators for functional response data used in the estimation of the reduced-form functions. Section 7 presents the empirical application. We provide notation, proofs of key results, and details about implementation of the methods in the empirical example in Appendices A–E. The Supplemental Material provides all remaining proofs, additional technical material, and results from a small Monte Carlo simulation (Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017)).

2. THE TREATMENT EFFECTS SETTING AND TARGET PARAMETERS

2.1. *Observables and Reduced-Form Parameters*

The observed random variables consist of $((Y_u)_{u\in\mathcal{U}}, X, Z, D)$. The outcome variable of interest Y_u is indexed by $u \in \mathcal{U}$. We give examples of the index u below. The variable $D \in \mathcal{D} = \{0, 1\}$ is a binary indicator of the receipt of a treatment or participation in a program. It will typically be treated as endogenous; that is, we will typically view the treatment as assigned non-randomly with respect to the outcome. The instrumental variable $Z \in \mathcal{Z} = \{0, 1\}$ is a binary indicator, such as an offer of participation, that is assumed to be randomly assigned conditional on the observable covariates X with support $\mathcal{X}$.[8] For example, we argue that 401(k) eligibility can be considered exogenous only after conditioning on income and other individual characteristics in the empirical application. The notions of exogeneity and endogeneity we employ are standard and thus omitted.[9]

The indexing of the outcome Y_u by u is useful to analyze functional data. For example, Y_u could represent an outcome falling short of a threshold, namely $Y_u = 1(Y \leq u)$, in the context of distributional analysis; Y_u could be a height indexed by age u in growth charts analysis; or 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 $\mathcal{U}$ to be a singleton set.

[8]Of course, by "randomly assigned" we mean independently of potential outcomes conditional on the covariates.

[9]For completeness, we provide a review of these conditions as well as restate standard conditions that are sufficient for a causal interpretation of the target parameters in the Supplemental Material.

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

$$\alpha_V(z) := \mathrm{E}_P\big[g_V(z, X)\big] \quad \text{and} \quad \gamma_V := \mathrm{E}_P[V], \tag{2.1}$$

where $z = 0$ or $z = 1$ are the fixed values of Z.[10] The function g_V maps $\mathcal{Z}\mathcal{X}$, the support of the vector (Z, X), to the real line $\mathbb{R}$ and is defined as

$$g_V(z, x) := \mathrm{E}_P[V|Z = z, X = x]. \tag{2.2}$$

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.

All 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 modern methods to deal with high-dimensional data coupled with orthogonal estimating equations. The orthogonality property allows us to deal 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 (2.1) are key because the structural parameters of interest are functionals of these elementary objects. The local average structural function (LASF) defined as

$$\theta_{Y_u}(d) = \frac{\alpha_{\mathbf{1}_d(D)Y_u}(1) - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\alpha_{\mathbf{1}_d(D)}(1) - \alpha_{\mathbf{1}_d(D)}(0)}, \quad d \in \{0, 1\} \tag{2.3}$$

underlies the formation of many commonly used treatment effects. Under standard assumptions, the LASF identifies average potential outcomes for the group of *compliers*, individuals whose treatment status may be influenced by variation in the instrument, in the treated and non-treated states; see, for example, Abadie (2002, 2003). The local average treatment effect (LATE) of Imbens and Angrist (1994) corresponds to the difference of the two values of the LASF:

$$\theta_{Y_u}(1) - \theta_{Y_u}(0). \tag{2.4}$$

The term local designates that this parameter does not measure the effect on the entire population but rather measures the effect on the subpopulation of compliers.[11]

When there is no endogeneity, formally when $D \equiv Z$, the LASF and LATE become the average structural function (ASF) and average treatment effect (ATE) on the entire

[10]The expectation that defines $\alpha_V(z)$ is well-defined under the standard support condition $0 < c < \mathrm{P}_P(Z = 1|X) < 1 - c$ a.s. This condition is standard in treatment effects estimation; see, for example, the Supplemental Material. We impose this condition in Assumption 4.1.

[11]The methods of the paper can be extended to analyze the marginal treatment effects of Heckman and Vytlacil (1999, 2005).

population. Thus, our results cover this situation as a special case where the ASF and ATE simplify to

$$\theta_{Y_u}(z) = \alpha_{Y_u}(z), \quad \theta_{Y_u}(1) - \theta_{Y_u}(0) = \alpha_{Y_u}(1) - \alpha_{Y_u}(0). \tag{2.5}$$

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

$$\alpha_{Y_u}(z), \quad \alpha_{Y_u}(1) - \alpha_{Y_u}(0).$$

Thus, the LASF and LATE are primary targets of interest in this paper, and the ASF and ATE are subsumed as special cases.

2.2.1. *Local Distribution and Quantile Treatment Effects*

Setting $Y_u = Y$ in (2.3) and (2.4) provides the conventional LASF and LATE. An important generalization arises by letting $Y_u = 1(Y \leq u)$ be the indicator of the outcome of interest falling below a threshold $u \in \mathbb{R}$. In this case, the family of effects

$$\big(\theta_{Y_u}(1) - \theta_{Y_u}(0)\big)_{u \in \mathbb{R}} \tag{2.6}$$

describe the local distribution treatment effects (LDTE). Similarly, we can look at the quantile left-inverse transform of the curve $u \longmapsto \theta_{Y_u}(d)$,

$$\theta_Y^{\leftarrow}(\tau, d) := \inf\big\{u \in \mathbb{R} : \theta_{Y_u}(d) \geq \tau\big\}, \tag{2.7}$$

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

$$\big(\theta_Y^{\leftarrow}(\tau, 1) - \theta_Y^{\leftarrow}(\tau, 0)\big)_{\tau \in (0,1)}. \tag{2.8}$$

The LQTE identify the differences of quantiles between the distribution of potential outcomes in the treated and non-treated states for compliers.

2.3. *Target Structural Parameters—Local Treatment Effects on the Treated*

We may also be interested in local treatment effects on the treated. The key object in defining these effects is the local average structural function on the treated (LASF-T) which is defined by its two values:

$$\vartheta_{Y_u}(d) = \frac{\gamma_{\mathbf{1}_d(D)Y_u} - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\gamma_{\mathbf{1}_d(D)} - \alpha_{\mathbf{1}_d(D)}(0)}, \quad d \in \{0, 1\}. \tag{2.9}$$

The LASF-T identifies average potential outcomes for the group of *treated compliers* in the treated and non-treated states under standard assumptions. The local average treatment effect on the treated (LATE-T) introduced in Hong and Nekipelov (2010) and Frölich and Melly (2013) is the difference of two values of the LASF-T:

$$\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0). \tag{2.10}$$

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 on the treated (ASF-T) and average treatment effect on the treated (ATE-T). In this special case, the ASF-T and ATE-T simplify to

$$\vartheta_{Y_u}(1) = \frac{\gamma_{\mathbf{1}_1(D)Y_u}}{\gamma_{\mathbf{1}_1(D)}}, \quad \vartheta_{Y_u}(0) = \frac{\gamma_{\mathbf{1}_0(D)Y_u} - \alpha_{Y_u}(0)}{\gamma_{\mathbf{1}_0(D)} - 1}, \quad \vartheta_{Y_u}(1) - \vartheta_{Y_u}(0), \tag{2.11}$$

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

2.3.1. *Local Distribution and Quantile Treatment Effects on the Treated*

Local distribution 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 $Y_u = 1(Y \leq u)$ be the indicator of the outcome of interest falling below a threshold u. The family of treatment effects

$$\big(\vartheta_{Y_u}(1) - \vartheta_{Y_u}(0)\big)_{u \in \mathbb{R}} \tag{2.12}$$

then describes the LDTE-T. We can also use the quantile left-inverse transform of the curve $u \longmapsto \vartheta_{Y_u}(d)$, namely $\vartheta_Y^{\leftarrow}(\tau, d) := \inf\{u \in \mathbb{R} : \vartheta_{Y_u}(d) \geq \tau\}$, and define the LQTE-T:

$$\big(\vartheta_Y^{\leftarrow}(\tau, 1) - \vartheta_Y^{\leftarrow}(\tau, 0)\big)_{\tau \in (0,1)}. \tag{2.13}$$

Under conditional exogeneity, LQTE and LQTE-T reduce to the quantile treatment effects (QTE) and quantile treatment effects on the treated (QTE-T) (Koenker (2005, Chapter 2)).

3. ESTIMATION OF REDUCED-FORM AND STRUCTURAL PARAMETERS IN A DATA-RICH ENVIRONMENT

The key objects used to define the structural parameters in Section 2 are the expectations

$$\alpha_V(z) = \mathrm{E}_P\big[g_V(z, X)\big] \quad \text{and} \quad \gamma_V = \mathrm{E}_P[V], \tag{3.1}$$

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 $\mathcal{V}_u$:

$$V \in \mathcal{V}_u := \{V_{uj}\}_{j=1}^5 := \big\{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\big\}. \tag{3.2}$$

It is clear that $g_V(z, X)$ will play an important role in estimating $\alpha_V(z)$. A related function that will also play an important role in forming a robust estimation strategy is the propensity score $m_Z : \mathcal{ZX} \longmapsto \mathbb{R}$ defined by

$$m_Z(z, x) := \mathrm{P}_P[Z = z|X = x]. \tag{3.3}$$

We will denote other potential values for the functions g_V and m_Z by the parameters g and m, respectively. We can then estimate $\alpha_V(z)$ by estimating g_V and m_Z using high-dimensional modeling and estimation methods.[12]

[12]Note that there is an alternative approach based on decomposing g_V as $g_V(z, x) = \sum_{d=0}^1 e_V(d, z, x) \times l_D(d, z, x)$ where the regression functions e_V and l_D map the support of (D, Z, X), $\mathcal{DZX}$, to the real line

In the rest of this section, we describe the estimation of the reduced-form and structural parameters. The estimation method consists of three steps:

(1) Estimate the predictive relationships m_Z and g_V using high-dimensional nonparametric methods with model selection.

(2) Estimate the reduced-form parameters α_V and γ_V using orthogonal estimating equations to immunize the reduced-form estimators to imperfect model selection in the first step.

(3) Estimate the structural parameters and effects via the plug-in rule.

3.1. *First Step: Modeling and Estimating g_V and m_Z*

In this section, we discuss estimation of the conditional expectation functions g_V and m_Z. Since these functions are unknown and potentially complicated, we use a generalized linear combination of a large number of control terms

$$f(X) = \big(f_j(X)\big)_{j=1}^{p}, \tag{3.4}$$

to approximate g_V and m_Z. Specifically, we use

$$g_V(z, x) =: \Lambda_V\big[f(z, x)'\beta_V\big] + r_V(z, x), \tag{3.5}$$

$$f(z, x) := \big((1-z)f(x)', zf(x)'\big)', \quad \beta_V := \big(\beta_V(0)', \beta_V(1)'\big)', \tag{3.6}$$

$$m_Z(1, x) =: \Lambda_Z\big[f(x)'\beta_Z\big] + r_Z(x), \tag{3.7}$$

$$m_Z(0, x) = 1 - \Lambda_Z\big[f(x)'\beta_Z\big] - r_Z(x).$$

In these equations, $r_V(z, x)$ and $r_Z(x)$ are approximation errors, and the functions $\Lambda_V(f(z, x)'\beta_V)$ 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 variable is 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 $\Lambda(u) = \Phi(u) = (2\pi)^{-1/2}\int_{-\infty}^{u} e^{-z^2/2}\,dz$ 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.

As discussed in the Introduction, the dictionary of controls, denoted by $f(X)$, can 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. We also note that the functions f 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 (3.5)–(3.7) that require only a small number of nonzero 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, for all $V \in \mathcal{V} := \{\mathcal{V}_u : u \in \mathcal{U}\}$,

$$\|\beta_V\|_0 + \|\beta_Z\|_0 \leq s, \tag{3.8}$$

and are defined by $e_V(d, z, x) := \mathrm{E}_P[V|D = d, Z = z, X = x]$ and $l_D(d, z, x) := \mathrm{P}_P[D = d|Z = z, X = x]$. We provide some discussion of this approach in the Supplemental Material.

where $\|x\|_0$ is the number of nonzero components of vector x and all other norms we use are defined in Appendix A. That is, there are at most $s = s_n \ll n$ components of $f(Z, X)$ and $f(X)$ with nonzero coefficient in the approximations to g_V and m_Z. 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 $V \in \mathcal{V}$,

$$\{\mathrm{E}_P[r_V^2(Z, X)]\}^{1/2} + \{\mathrm{E}_P[r_Z^2(X)]\}^{1/2} \lesssim \sqrt{s/n}. \tag{3.9}$$

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

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

These conditions ensure that the functions g_V and m_Z are estimable at a $o(n^{-1/4})$ rate and are used to derive asymptotic normality results for the structural and reduced-form parameter estimators. They could be relaxed through the use of sample splitting methods as in Belloni et al. (2012).

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 small relative to the sample size.[13] 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 and number 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 methods 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 the 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).

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

$$\hat{\beta}_V \in \arg \min_{\beta \in \mathbb{R}^{\dim(\tilde{X})}} \left(\mathbb{E}_n[M(\tilde{Y}, \tilde{X}'\beta)] + \frac{\lambda}{n} \|\hat{\Psi}\beta\|_1 \right), \tag{3.10}$$

where $\hat{\Psi} = \mathrm{diag}(\hat{l}_1, \ldots, \hat{l}_{\dim(\tilde{X})})$ is a diagonal matrix of data-dependent penalty loadings, $M(y, t) = (y - t)^2/2$ in the case of linear regression, and $M(y, t) = -\{1(y = 1) \log \Lambda_V(t) + 1(y = 0) \log(1 - \Lambda_V(t))\}$ in the case of binary regression. The penalty level, λ, and loadings, $\hat{l}_j$, $j = 1, \ldots, \dim(\tilde{X})$, are selected to guarantee good theoretical properties of the

[13]For example, one would select a set of basis functions, $\{f_j(X)\}_{j=1}^{\infty}$, such as power series or splines, and then use only the first $s \ll n$ terms in the basis under the assumption that $s^C/n \to 0$ for some number C whose value depends on the specific context in a standard nonparametric approach using series.

method. We provide further discussion of these methods for estimation of a continuum of functions in Section 6, and we specify detailed implementation algorithms used in the empirical example in Appendix F. 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 $\hat{\beta}_V$ solely as a model selection device. Specifically, it makes use of the labels of the regressors with nonzero estimated coefficients, $\hat{I}_V := \operatorname{supp}(\hat{\beta}_V)$. The Post-Lasso estimator is then a solution to

$$\tilde{\beta}_V \in \arg \min_{\beta \in \mathbb{R}^{\dim(\tilde{X})}} \left(\mathbb{E}_n\left[M\left(\tilde{Y}, \tilde{X}'\beta\right)\right] : \beta_j = 0, j \notin \hat{I}_V\right). \tag{3.11}$$

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

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

3.2. *Second Step: Robust Estimation of the Reduced-Form Parameters $\alpha_V(z)$ and γ_V*

Estimation of the key quantities $\alpha_V(z)$ will make heavy use of orthogonal moment functions as defined in (1.2). These moment functions are closely tied to efficient influence functions, where efficiency is in the sense of locally minimax semiparametric efficiency. The use of these functions will deliver robustness with respect to the non-regularity of the post-selection and penalized estimators needed to manage high-dimensional data. The use of these functions also automatically delivers semiparametric efficiency for estimating and performing inference on the reduced-form parameters and their smooth transformations—the structural parameters.

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

$$\psi^\alpha_{V,z}(W) := \psi^\alpha_{V,z,g_V,m_Z}\left(W, \alpha_V(z)\right) \quad \text{and} \tag{3.12}$$

$$\psi^\alpha_{V,z,g,m}(W, \alpha) := \frac{1(Z = z)\left(V - g(z, X)\right)}{m(z, X)} + g(z, X) - \alpha. \tag{3.13}$$

This efficient influence function was derived by Hahn (1998); it has recently been used by Cattaneo (2010) in the series context (with $p \ll n$) and Rothe and Firpo (2013) in the kernel context. The efficient influence function and the moment function for γ_V are trivially given by

$$\psi^\gamma_V(W) := \psi^\gamma_V(W, \gamma_V) \quad \text{and} \quad \psi^\gamma_V(W, \gamma) := V - \gamma. \tag{3.14}$$

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

$$\mathbb{E}_n\big[\psi^{\alpha}_{V,z,\hat{g}_V,\hat{m}_Z}(W,\alpha)\big] = 0, \quad \mathbb{E}_n\big[\psi^{\gamma}_V(W,\gamma)\big] = 0, \tag{3.15}$$

where $\hat{g}_V$ and $\hat{m}_Z$ are constructed as in Section 3.1. We apply this procedure to each variable name $V \in \mathcal{V}_u$ and obtain the estimator[14]

$$\hat{\rho}_u := \big(\big\{\hat{\alpha}_V(0), \hat{\alpha}_V(1), \hat{\gamma}_V\big\}\big)_{V\in\mathcal{V}_u} \quad \text{of} \quad \rho_u := \big(\big\{\alpha_V(0), \alpha_V(1), \gamma_V\big\}\big)_{V\in\mathcal{V}_u}. \tag{3.16}$$

The estimator and the parameter are vectors in $\mathbb{R}^{d_\rho}$ with dimension $d_\rho = 3 \times \dim \mathcal{V}_u = 15$.

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

$$\begin{aligned} &\sqrt{n}(\hat{\rho}_u - \rho_u) \rightsquigarrow N\big(0, \mathrm{Var}_P\big(\psi^{\rho}_u\big)\big), \\ &\psi^{\rho}_u := \big(\big\{\psi^{\alpha}_{V,0}, \psi^{\alpha}_{V,1}, \psi^{\gamma}_V\big\}\big)_{V\in\mathcal{V}_u}, \quad \text{uniformly in } P \in \mathcal{P}_n, \end{aligned} \tag{3.17}$$

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

We then stack all the reduced-form estimators and parameters over $u \in \mathcal{U}$ as

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

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

$$\sqrt{n}(\hat{\rho} - \rho) \rightsquigarrow Z_P := \big(\mathbb{G}_P\psi^{\rho}_u\big)_{u\in\mathcal{U}}, \quad \text{uniformly in } P \in \mathcal{P}_n, \tag{3.18}$$

where $\mathbb{G}_P$ denotes the P-Brownian bridge (van der Vaart and Wellner (1996, pp. 81–82)). This result contains (3.17) 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 verifies that the functional central limit theorem applies to 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 a multiplier bootstrap method as in Giné and Zinn (1984). We define $\hat{\rho}^* = (\hat{\rho}^*_u)_{u\in\mathcal{U}}$, a bootstrap draw of $\hat{\rho}$, via

$$\hat{\rho}^*_u = \hat{\rho}_u + n^{-1}\sum_{i=1}^{n} \xi_i \hat{\psi}^{\rho}_u(W_i). \tag{3.19}$$

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 P_ξ does not depend on P. We also impose that

$$\mathrm{E}[\xi] = 0, \quad \mathrm{E}\big[\xi^2\big] = 1, \quad \mathrm{E}\big[\exp\big(|\xi|\big)\big] < \infty. \tag{3.20}$$

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

Examples of ξ include (a) $\xi = \mathcal{E} - 1$, where $\mathcal{E}$ is a standard exponential random variable, (b) $\xi = \mathcal{N}$, where $\mathcal{N}$ is a standard normal random variable, and (c) $\xi = \mathcal{N}_1/\sqrt{2} + (\mathcal{N}_2^2 - 1)/2$, where $\mathcal{N}_1$ and $\mathcal{N}_2$ are mutually independent standard normal random variables.[15] The choices of (a), (b), and (c) correspond respectively to the Bayesian bootstrap (e.g., Hahn (1997) and Chamberlain and Imbens (2003)), the Gaussian multiplier method (e.g, Giné and Zinn (1984) and van der Vaart and Wellner (1996, Chapter 3.6)), and the wild bootstrap method (Mammen (1993)).[16] $\hat{\psi}_u^\rho$ in (3.19) is an estimator of the influence function ψ_u^ρ defined via the plug-in rule:

$$
\begin{aligned}
(3.21)\qquad & \hat{\psi}_u^\rho = \big(\hat{\psi}_V^\rho\big)_{V \in \mathcal{V}_u}, \\
& \hat{\psi}_V^\rho(W) := \big\{\psi^\alpha_{V,0,\hat{g}_V,\hat{m}_Z}\big(W, \hat{\alpha}_V(0)\big), \psi^\alpha_{V,1,\hat{g}_V,\hat{m}_Z}\big(W, \hat{\alpha}_V(1)\big), \psi^\gamma_V(W, \hat{\gamma}_V)\big\}.
\end{aligned}
$$

Note that this bootstrap is computationally efficient since it does not involve recomputing the influence functions $\hat{\psi}_u^\rho$.[17] Each new draw of $(\xi_i)_{i=1}^n$ generates a new draw of $\hat{\rho}^*$ holding the data and the estimates of the influence functions fixed. This method simply amounts to resampling the first-order approximations to the estimators. Here we build upon prior uses of this or similar methods in low-dimensional settings such as Hansen (1996) and Kline and Santos (2012).

We establish that the bootstrap law of $\sqrt{n}(\hat{\rho}^* - \hat{\rho})$ is uniformly asymptotically consistent: In the metric space $\ell^\infty(\mathcal{U})^{d_\rho}$, conditionally on the data,

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

where $\rightsquigarrow_B$ denotes weak convergence of the bootstrap law in probability, as defined in Appendix B.

3.3. *Third Step: Robust Estimation of the Structural Parameters*

All structural parameters we consider take the form of smooth transformations of the reduced-form parameters:

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

The structural parameters may themselves carry an index $q \in \mathcal{Q}$ that can be different from u; for example, the LQTE is indexed by a quantile index $q \in (0, 1)$. This formulation includes as special cases all the structural functions of Section 2. We estimate these quantities by the plug-in rule. We establish the 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 (extended to handle uniformity in P).

For the application of the functional delta method, we require that the functional $\rho \longmapsto \phi(\rho)$ be Hadamard differentiable *uniformly* in $\rho \in \mathbb{D}_\rho$, where $\mathbb{D}_\rho$ is a set that contains the true values $\rho = \rho_P$ for all $P \in \mathcal{P}_n$, tangentially to a subset that contains the realizations

[15]We do not consider the nonparametric bootstrap, which corresponds to using multinomial multipliers ξ, to reduce the length of the paper; but we note that the conditions and analysis could be extended to cover this case.

[16]The motivation for method (c) is that it is able to match three 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.

[17]Chernozhukov and Hansen (2006) and Hong and Scaillet (2006) proposed related computationally efficient bootstrap schemes that resample the influence functions.

of Z_P for all $P \in \mathcal{P}_n$ with derivative map $h \longmapsto \phi'_\rho(h) = (\phi'_\rho(h)(q))_{q\in\mathcal{Q}}$.[18] We define the estimators of the structural parameters and their bootstrap versions via the plug-in rule as

$$(3.23)\qquad \hat{\Delta} := (\hat{\Delta}_q)_{q\in\mathcal{Q}}, \quad \hat{\Delta}_q := \phi(\hat{\rho})(q), \quad \text{and}$$
$$\hat{\Delta}^* := \big(\hat{\Delta}^*_q\big)_{q\in\mathcal{Q}}, \quad \hat{\Delta}^*_q := \phi\big(\hat{\rho}^*\big)(q).$$

We establish that these estimators are asymptotically Gaussian:

$$(3.24)\qquad \sqrt{n}(\hat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P), \quad \text{uniformly in } P \in \mathcal{P}_n,$$

and that the bootstrap consistently estimates their large sample distribution:

$$(3.25)\qquad \sqrt{n}\big(\hat{\Delta}^* - \hat{\Delta}\big) \rightsquigarrow_B \phi'_\rho(Z_P), \quad \text{uniformly in } P \in \mathcal{P}_n.$$

These results can be used to construct simultaneous confidence bands and test functional hypotheses on Δ using the methods described, for example, in Chernozhukov and Fernández-Val (2005) and Chernozhukov, Fernández-Val, and Melly (2013).

4. THEORY: ESTIMATION AND INFERENCE ON LOCAL TREATMENT EFFECTS FUNCTIONALS

Consider fixed sequences of numbers $\delta_n \searrow 0$, $\epsilon_n \searrow 0$, $\Delta_n \searrow 0$, at a speed at most polynomial in n (e.g., $\delta_n \geq 1/n^c$ for some $c > 0$), $\ell_n := \log n$, and positive constants c, C, and $c' < 1/2$. These sequences and constants will not vary with P. The probability P can vary in the set $\mathcal{P}_n$ of probability measures, termed "data-generating processes," where $\mathcal{P}_n$ is typically a set that is weakly increasing in n, that is, $\mathcal{P}_n \subseteq \mathcal{P}_{n+1}$. Other definitions and notation are collected in Appendix A.

ASSUMPTION 4.1—Basic Assumptions: (i) *Consider a random element W with values in a measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ and law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((W_{ui})_{u\in\mathcal{U}})_{i=1}^n$ consist of n i.i.d. copies of a random element $(W_u)_{u\in\mathcal{U}} = ((Y_u)_{u\in\mathcal{U}}, D, Z, X)$, where $\mathcal{U}$ is a Polish space equipped with its Borel sigma-field and $(Y_u, D, Z, X) \in \mathbb{R}^{3+d_X}$. Each W_u is generated via a measurable transform $t(W, u)$ of W and u, namely the map $t : \mathcal{W} \times \mathcal{U} \longmapsto \mathbb{R}^{3+d_X}$ is measurable, and the map can possibly depend on P. Let*

$$\mathcal{V}_u := \{V_{uj}\}_{j\in\mathcal{J}} := \big\{Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\big\},$$
$$\mathcal{V} := (\mathcal{V}_u)_{u\in\mathcal{U}},$$

where $\mathcal{J} = \{1, \dots, 5\}$. (ii) *For $\mathcal{P} := \bigcup_{n=n_0}^\infty \mathcal{P}_n$, the map $u \longmapsto Y_u$ obeys the uniform continuity property*:

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

where the second supremum in the first expression is taken over $u, \bar{u} \in \mathcal{U}$, and $\mathcal{U}$ is a totally bounded metric space equipped with a semimetric $d_\mathcal{U}$. The uniform covering entropy of the

[18]We give the definition of uniform Hadamard differentiability in Definition B.1 of Appendix B.

set $\mathcal{F}_P = \{Y_u : u \in \mathcal{U}\}$, *viewed as a collection of maps* $(\mathcal{W}, \mathcal{A}_{\mathcal{W}}) \longmapsto \mathbb{R}$, *obeys*

$$\sup_Q \log N\big(\epsilon \|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}\big) \le C \log(\mathrm{e}/\epsilon) \vee 0$$

for all $P \in \mathcal{P}$, *where* $F_P(W) = \sup_{u\in\mathcal{U}} |Y_u|$, *with the supremum taken over all finitely discrete probability measures* Q *on* $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$. (iii) *For each* $P \in \mathcal{P}$, *the conditional probability of* $Z = 1$ *given* X *is bounded away from zero or* 1, *namely* $c' \le m_Z(1, X) \le 1 - c'$ *P-a.s., the instrument* Z *has a nontrivial impact on* D, *namely* $c' \le |\mathrm{P}_P[D = 1|Z = 1, X] - \mathrm{P}_P[D = 1|Z = 0, X]|$ *P-a.s., and the regression function* g_V *is bounded*, $\|g_V\|_{P,\infty} < \infty$ *for all* $V \in \mathcal{V}$.

Assumption 4.1 is stated to deal with the measurability issues associated with functional response data. This assumption also implies that the set of functions $(\psi_u^\rho)_{u\in\mathcal{U}}$, where

$$\psi_u^\rho := \big(\big\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\big\}\big)_{V\in\mathcal{V}_u},$$

is P-Donsker uniformly in $\mathcal{P}$. That is, it implies

$$Z_{n,P} \rightsquigarrow Z_P \quad \text{in} \quad \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}, \tag{4.1}$$

$$Z_{n,P} := \big(\mathbb{G}_n \psi_u^\rho\big)_{u\in\mathcal{U}} \quad \text{and} \quad Z_P := \big(\mathbb{G}_P \psi_u^\rho\big)_{u\in\mathcal{U}}, \tag{4.2}$$

with $\mathbb{G}_P$ denoting the P-Brownian bridge (van der Vaart and Wellner (1996, pp. 81–82)) and with Z_P having bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{u\in\mathcal{U}} \big\|Z_P(u)\big\| < \infty, \quad \lim_{\varepsilon\searrow 0} \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u,\tilde{u})\le\varepsilon} \big\|Z_P(u) - Z_P(\tilde{u})\big\| = 0. \tag{4.3}$$

We work with the sequence of constants defined prior to Assumption 4.1.

ASSUMPTION 4.2—Approximate Sparsity: *Under each* $P \in \mathcal{P}_n$ *and for each* $n \ge n_0$, *uniformly for all* $V \in \mathcal{V}$: (i) *The approximations* (3.5)–(3.7) *hold with the link functions* Λ_V *and* Λ_Z *belonging to the set* $\mathcal{L}$, *the sparsity condition* $\|\beta_V\|_0 + \|\beta_Z\|_0 \le s$ *holding, the approximation errors satisfying* $\|r_V\|_{P,2} + \|r_Z\|_{P,2} \le \delta_n n^{-1/4}$ *and* $\|r_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \le \epsilon_n$, *and the sparsity index* s *and the number of terms* p *in the vector* $f(X)$ *obeying* $s^2 \log^2(p \vee n) \log^2 n \le \delta_n n$. (ii) *There are estimators* $\bar{\beta}_V$ *and* $\bar{\beta}_Z$ *such that, with probability no less than* $1 - \Delta_n$, *the estimation errors satisfy* $\|f(Z, X)'(\bar{\beta}_V - \beta_V)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z - \beta_Z)\|_{\mathbb{P}_n,2} \le \delta_n n^{-1/4}$, $K_n\|\bar{\beta}_V - \beta_V\|_1 + K_n\|\bar{\beta}_Z - \beta_Z\|_1 \le \epsilon_n$; *the estimators are sparse such that* $\|\bar{\beta}_V\|_0 + \|\bar{\beta}_Z\|_0 \le Cs$; *and the empirical and population norms induced by the Gram matrix formed by* $(f(X_i))_{i=1}^n$ *are equivalent on sparse subsets*, $\sup_{\|\delta\|_0\le\ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2}/\|f(X)'\delta\|_{P,2} - 1| \le \epsilon_n$. (iii) *The following boundedness conditions hold*: $\|\|f(X)\|_\infty\|_{P,\infty} \le K_n$ *and* $\|V\|_{P,\infty} \le C$.

COMMENT 4.1: Assumption 4.2 imposes simple intermediate-level conditions which encode both the approximate sparsity of the models as well as some reasonable behavior of the sparse estimators of m_Z and g_V. These conditions significantly extend and generalize the conditions employed in the literature on adaptive estimation using series methods. 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. Sufficient conditions for the equivalence between empirical and population norms and primitive examples of functions admitting sparse approximations are given in Belloni, Chernozhukov, and Hansen (2014). We provide primitive conditions for Lasso estimators to satisfy the bounds above while addressing the problem of estimating continua of approximately sparse nuisance functions in Section 6. We expect that other sparsity-based estimators, such as the Dantzig selector or adaptive Lasso, could be used in the present context as well.

Under the stated assumptions, the empirical reduced-form process $\hat{Z}_{n,P} = \sqrt{n}(\hat{\rho} - \rho)$ defined by (3.16) obeys the following relations. We recall definitions of convergence uniformly in $P \in \mathcal{P}_n$ in Appendix A.

THEOREM 4.1—Uniform Gaussianity of the Reduced-Form Parameter Process: *Under Assumptions 4.1 and 4.2, the reduced-form empirical process admits a linearization*; *namely*,

$$\hat{Z}_{n,P} := \sqrt{n}(\hat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in} \quad \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.4}$$

The process $\hat{Z}_{n,P}$ is asymptotically Gaussian, *namely*,

$$\hat{Z}_{n,P} \rightsquigarrow Z_P \quad \text{in} \quad \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n, \tag{4.5}$$

where Z_P is defined in (4.2) *and its paths obey the property* (4.3).

Another main result of this section shows that the bootstrap law of the process

$$\hat{Z}^*_{n,P} := \sqrt{n}\big(\hat{\rho}^* - \hat{\rho}\big) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \hat{\psi}^\rho_u(W_i),$$

where $\hat{\psi}^\rho_u$ is defined in (3.21), provides a valid approximation to the large sample law of $\sqrt{n}(\hat{\rho} - \rho)$.

THEOREM 4.2—Validity of Multiplier Bootstrap for Inference on Reduced-Form Parameters: *Under Assumptions 4.1 and 4.2, the bootstrap law consistently approximates the large sample law Z_P of $Z_{n,P}$ uniformly in $P \in \mathcal{P}_n$, namely,*

$$\hat{Z}^*_{n,P} \rightsquigarrow_B Z_P \quad \text{in} \quad \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.6}$$

Next we consider inference on the structural functionals Δ defined in (3.22). We derive the large sample distribution of the estimator $\hat{\Delta}$ in (3.23), and show that the multiplier bootstrap law of $\hat{\Delta}^*$ in (3.23) provides a consistent approximation to that distribution. We rely on the functional delta method in our derivations, which we modify to handle uniformity with respect to the underlying d.g.p. P. Our argument relies on the following assumption on the structural functionals.

ASSUMPTION 4.3—Uniform Hadamard Differentiability of Structural Functionals: *Suppose that for each $P \in \mathcal{P}$, $\rho = \rho_P \in \mathbb{D}_\rho$, a compact metric space. Suppose $\varrho \longmapsto \phi(\varrho)$, a functional of interest mapping $\mathbb{D}_\phi \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}$ to $\ell^\infty(\mathcal{Q})$, where $\mathbb{D}_\rho \subset \mathbb{D}_\phi$, is Hadamard differentiable in ϱ tangentially to $\mathbb{D}_0 = UC(\mathcal{U})^{d_\rho}$ uniformly in $\varrho \in \mathbb{D}_\rho$, with the linear derivative map $\phi'_\varrho : \mathbb{D}_0 \longmapsto \mathbb{D}$ such that the mapping $(\varrho, h) \longmapsto \phi'_\varrho(h)$ from $\mathbb{D}_\rho \times \mathbb{D}_0$ to $\ell^\infty(\mathcal{Q})$ is continuous.*

The definition of uniform Hadamard differentiability is given in Definition B.1 of Appendix B. Assumption 4.3 holds for all examples of structural parameters listed in Section 2.

The following corollary gives the large sample law of $\sqrt{n}(\hat{\Delta} - \Delta)$, the properly normalized structural estimator. It also shows that the bootstrap law of $\sqrt{n}(\hat{\Delta}^* - \hat{\Delta})$, computed conditionally on the data, approaches the large sample law $\sqrt{n}(\hat{\Delta} - \Delta)$. It follows from the previous theorems as well as from a more general result contained in Theorem 5.3.

COROLLARY 4.1—Limit Theory and Validity of Multiplier Bootstrap for Smooth Structural Functionals: *Under Assumptions* 4.1, 4.2, *and* 4.3,

$$\sqrt{n}(\hat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\rho_P}(Z_P) \quad \text{in} \quad \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n, \tag{4.7}$$

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

$$\sqrt{n}\big(\hat{\Delta}^* - \hat{\Delta}\big) \rightsquigarrow_B T_P \quad \text{in} \quad \ell^\infty(\mathcal{Q}), \text{ uniformly in } P \in \mathcal{P}_n. \tag{4.8}$$

5. GENERAL THEORY: HONEST INFERENCE IN GENERAL MOMENT-CONDITION PROBLEMS WITH NUISANCE FUNCTIONS ESTIMATED BY MACHINE LEARNING METHODS

In this section, we consider a general moment-condition framework, where possibly a continuum of target parameters is of interest, and we use modern machine learning methods, with Lasso-type methods being a lead example, to estimate a continuum of high-dimensional nuisance functions. This setting covers a rich variety of modern moment-condition problems in econometrics including the treatment effects problem. We establish a functional central limit theorem for the estimators of the continuum of target parameters that holds uniformly in $P \in \mathcal{P}$, where $\mathcal{P}$ includes a wide range of data-generating processes with well-approximable continuums of nuisance functions. We also derive a functional central limit theorem for the multiplier bootstrap that resamples the first-order approximations to the standardized estimators of the continuum of target parameters and establish its uniform validity. Moreover, we establish the uniform validity of the functional delta method and the functional delta method for the multiplier bootstrap for smooth functionals of the continuum of target parameters using an appropriate strengthening of Hadamard differentiability.

5.1. *Setting*

We are interested in function-valued target parameters indexed by $u \in \mathcal{U} \subset \mathbb{R}^{d_u}$. We denote the true value of the target parameter by

$$\theta^0 = (\theta_u)_{u \in \mathcal{U}}, \quad \text{where} \quad \theta_u \in \Theta_u \subset \Theta \subset \mathbb{R}^{d_\theta}, \text{ for each } u \in \mathcal{U}.$$

We assume that for each $u \in \mathcal{U}$, the true value θ_u is identified as the solution to the following moment condition:

$$\mathrm{E}_P\big[\psi_u\big(W_u, \theta_u, h_u(Z_u)\big)\big] = 0, \tag{5.1}$$

where W_u is a random vector that takes values in a Borel set $\mathcal{W}_u \subset \mathbb{R}^{d_w}$ and contains as a subcomponent the vector Z_u taking values in a Borel set $\mathcal{Z}_u$, the moment function

$$\begin{aligned} &\psi_u : \mathcal{W}_u \times \Theta_u \times T_u \longmapsto \mathbb{R}^{d_\theta}, \\ &(w, \theta, t) \longmapsto \psi_u(w, \theta, t) = \big(\psi_{uj}(w, \theta, t)\big)_{j=1}^{d_\theta} \end{aligned} \tag{5.2}$$

is a Borel measurable map, and the function

$$h_u : \mathcal{Z}_u \longmapsto \mathbb{R}^{d_t}, \quad z \longmapsto h_u(z) = \big(h_{um}(z)\big)_{m=1}^{d_t} \in T_u(z), \tag{5.3}$$

is another Borel measurable map that denotes the possibly infinite-dimensional nuisance parameter. The sets $T_u(z)$ are assumed to be convex for each $u \in \mathcal{U}$ and $z \in \mathcal{Z}_u$. Finite-dimensional nuisance parameters that do not depend on Z_u are treated as part of h_u as well.

We assume that the continuum of nuisance functions $(h_u)_{u\in\mathcal{U}}$ is well-approximable and can be well estimated by the modern generation of statistical and machine learning methods. In particular, our regularity conditions allow for approximately sparse nuisance functions, which can be modeled and estimated using methods such as Lasso and Post-Lasso. We let $\hat{h}_u = (\hat{h}_{um})_{m=1}^{d_t}$ denote the estimator of h_u, which we assume obeys the conditions in Assumption 5.3. The estimator $\hat{\theta}_u$ of θ_u is constructed as any approximate ϵ_n-solution in Θ_u to a sample analog of the moment condition (5.1), that is,

$$\begin{aligned} &\big\|\mathbb{E}_n\big[\psi_u\big(W_u, \hat{\theta}_u, \hat{h}_u(Z_u)\big)\big]\big\| \\ &\quad \le \inf_{\theta\in\Theta_u} \big\|\mathbb{E}_n\big[\psi_u\big(W_u, \theta, \hat{h}_u(Z_u)\big)\big]\big\| + \epsilon_n, \quad \text{where} \quad \epsilon_n = o\big(n^{-1/2}\big). \end{aligned} \tag{5.4}$$

COMMENT 5.1—Handling Over-Identified Cases: We do not analyze over-identified cases explicitly, but it is helpful to note that they can be handled within the current framework. Let $\psi_u^o(W_u, \theta, h_u^o(Z_u))$ be the original over-identifying moment function. Let $A_u(Z_u)$ denote the pointwise optimal matrix of linear combinations of the moments, so that the final moment function $\psi_u(W_u, \theta, h(Z_u)) = A_u(Z_u)\psi_u^o(W_u, \theta, h_u^o(Z_u))$ has the same dimension as θ_u. Here $h_u(Z_u) = (\mathrm{vec}(A_u(Z_u))', h_u^{o'}(Z_u))'$; that is, we simply treat A_u as part of the nuisance function h_u being estimated. We do not analyze the preliminary estimation of A_u in the present paper in order to maintain the focus on exactly identified cases as in Section 4.

5.2. *The Neyman Orthogonality or Immunization Condition*

A key condition needed for regular estimation of θ_u is an orthogonality or immunization condition. The simplest to explain, yet strongest, form of this condition can be expressed as follows:

$$\partial_t \mathrm{E}_P\big[\psi_u(W_u, \theta_u, t)|Z_u\big]|_{t=h_u(Z_u)} = 0, \quad \text{a.s.,} \tag{5.5}$$

subject to additional technical conditions such as continuity (5.6) and dominance (5.7) stated below, where we use the symbol ∂_t to abbreviate $\frac{\partial}{\partial t'}$. This condition holds in the previous setting of inference on treatment effects after interchanging the order of the derivative and expectation. The formulation here also covers certain non-smooth cases such as structural and instrumental quantile regression problems.

In the formal development, we use a more general form of the orthogonality condition.

DEFINITION 5.1—Neyman Orthogonality for Moment-Condition Models, General Form: For each $u \in \mathcal{U}$, suppose that (5.1)–(5.3) hold. Consider $\mathcal{H}_u$, a set of measurable functions $z \longmapsto h(z) \in T_u(z)$ from $\mathcal{Z}_u$ to $\mathbb{R}^{d_t}$ such that $\|h(Z_u) - h_u(Z_u)\|_{P,2} < \infty$ for all $h \in \mathcal{H}_u$. Suppose also that the set $T_u(z)$ is a convex subset of $\mathbb{R}^{d_t}$ for each $z \in \mathcal{Z}_u$. We say that ψ_u obeys a general form of orthogonality with respect to $\mathcal{H}_u$ uniformly in $u \in \mathcal{U}$ if the following conditions hold: For each $u \in \mathcal{U}$, the derivative

$$t \longmapsto \partial_t \mathrm{E}_P\big[\psi_u(W_u, \theta_u, t)|Z_u\big] \quad \text{is continuous on } t \in T_u(Z_u)\ P\text{-a.s.}; \tag{5.6}$$

is dominated,

$$\Big\| \sup_{t \in T_u(Z_u)} \big\|\partial_t \mathrm{E}_P\big[\psi_u(W_u, \theta_u, t)|Z_u\big]\big\| \Big\|_{P,2} < \infty; \tag{5.7}$$

and obeys the orthogonality condition:

$$\mathrm{E}_P\big[\partial_t \mathrm{E}_P\big[\psi_u\big(W_u, \theta_u, h_u(Z_u)\big)|Z_u\big]\big(h(Z_u) - h_u(Z_u)\big)\big] = 0 \quad \text{for all } h \in \mathcal{H}_u. \tag{5.8}$$

The orthogonality condition (5.8) reduces to (5.5) when $\mathcal{H}_u$ can span all measurable functions $h : \mathcal{Z}_u \longmapsto T_u$ such that $\|h\|_{P,2} < \infty$, but is more general otherwise.

COMMENT 5.2—An Alternative Formulation of the Neyman Orthogonality Condition: A slightly more general, though less primitive, definition of the orthogonality condition is as follows. For each $u \in \mathcal{U}$, suppose that (5.1)–(5.3) hold. Consider $\mathcal{H}_u$, a set of measurable functions $z \longmapsto h(z) \in T_u(z)$ from $\mathcal{Z}_u$ to $\mathbb{R}^{d_t}$ such that $\|h(Z_u) - h_u(Z_u)\|_{P,2} < \infty$ for all $h \in \mathcal{H}_u$, where the set $T_u(z)$ is a convex subset of $\mathbb{R}^{d_t}$ for each $z \in \mathcal{Z}_u$. We say that ψ_u obeys a general form of orthogonality with respect to $\mathcal{H}_u$ uniformly in $u \in \mathcal{U}$, if the following conditions hold: The Gateaux derivative map

$$\mathrm{D}_{u,t}[h - h_u] := \partial_t \mathrm{E}_P\big(\psi_u\big\{W_u, \theta_u, h_u(Z_u) + t\big[h(Z_u) - h_u(Z_u)\big]\big\}\big)$$

exists for all $t \in [0, 1)$, $h \in \mathcal{H}_u$, and $u \in \mathcal{U}$ and vanishes at $t = 0$—namely,

$$\mathrm{D}_{u,0}[h - h_u] = 0 \quad \text{for all } h \in \mathcal{H}_u. \tag{5.9}$$

Definition 5.1 implies this definition by the mean-value expansion and the dominated convergence theorem.

COMMENT 5.3—Orthogonalization Typically Expands the Number of Nuisance Parameters: It is important to use a moment function ψ_u that satisfies the orthogonality property given in (5.8); see examples given below. Generally, if we have a moment function $\tilde{\psi}_u$ which identifies θ_u but does not have this property, we can construct a moment function ψ_u that identifies θ_u and has the required orthogonality property by projecting the original function $\tilde{\psi}_u$ onto the orthocomplement of the tangent space for the original set of nuisance functions h_u^o; see, for example, van der Vaart and Wellner (1996), van der Vaart (1998, Chapter 25), Kosorok (2008), Belloni, Chernozhukov, and Kato (2013), and Belloni, Chernozhukov, and Hansen (2014). This projection creates the semiparametrically efficient score function. There are other ways to create orthogonal nuisance functions, as illustrated by the second example below.

Note that the projection typically depends on P, which gives rise to additional nuisance parameters h_u^n, which are then incorporated together with the original nuisance parameters into the new parameter $h_u = (h_u^0, h_u^n)$. Note that this is a feature of all of the examples we consider. For example, the orthogonal moment functions in the exogenous case of the treatment effects framework depend on both the regression function and the propensity score function. This point is clarified further by considering the classical linear model as demonstrated in the next remark.

EXAMPLE 1—Neyman Orthogonal Equations for Linear Regression: To illustrate the orthogonality condition in the simplest possible setting, let us consider the linear model:

$$Y = D\theta_0 + X'\beta_0 + \epsilon, \quad \mathrm{E}_P[\epsilon X] = 0, \quad D = X'\pi_0 + \nu, \quad \mathrm{E}_P[\nu X] = 0, \tag{5.10}$$

where D is the treatment and X are the controls of high dimension $p \gg n$. Call the first equation the regression equation, and the second equation the propensity score equation. The orthogonal moment condition that identifies the projection coefficient θ_0 is the Frisch–Waugh–Lovell partialling out interpretation of θ_0:

$$\mathrm{E}_P(U - \nu\theta_0)\nu = 0, \tag{5.11}$$

where U is the population residual left after projecting out the controls X from the outcome, that is, $Y = X'\delta_0 + U$, $\mathrm{E}_P UX = 0$; and ν is the population residual left after projecting out controls from the treatment as defined in the propensity score equation. The high-dimensional nuisance function is $h(Z) = (X'\delta, X'\pi)'$, for $Z = X$, with true value denoted by $h_0(Z) = (X'\delta_0, X'\pi_0)'$. Now the moment function

$$\psi\big(W, \theta, h(Z)\big) = \big\{\big(Y - X'\delta\big) - \big(D - X'\pi\big)\theta\big\}\big(D - X'\pi\big) \tag{5.12}$$

has the required orthogonality property (5.8), since by the law of iterated expectations and some simple algebra:

$$\begin{aligned} &\mathrm{E}_P\big[\partial_t \mathrm{E}_P\big[\psi\big(W, \theta_0, h_0(Z)\big) \mid Z\big]\big(h(Z) - h_0(Z)\big)\big] \\ &\quad = \big(\mathrm{E}_P\big[-\big(D - X'\pi_0\big)X'a\big], \mathrm{E}_P\big[\big\{-\big(Y - X'\delta_0\big) + 2\big(D - X'\pi_0\big)\theta_0\big\}X'b\big]\big) \\ &\quad = 0, \end{aligned} \tag{5.13}$$

for $a = \delta - \delta_0$ and $b = \pi - \pi_0$. In fact, $\psi(W, \theta_0, h_0)$ is the semiparametrically efficient score for θ_0. The resulting estimator of θ_0 is root-n consistent and asymptotically normal, uniformly within a class of approximately sparse models as follows from the general results of this section, and is also semiparametrically efficient. See also Belloni, Chernozhukov, and Hansen (2014) which deals with the partially linear model in detail and thus covers this linear example as a special case.

Note that the orthogonal moment function contains two nuisance functions: the regression function and the propensity score—$X'\delta_0$ and $X'\pi_0$. We could also identify θ_0 through non-orthogonal moment conditions containing single nuisance functions:

$$\mathrm{E}_P\big[\big\{Y - D\theta_0 - X'\beta_0\big\}D\big] = 0 \quad \text{or} \quad \mathrm{E}_P\big[\{Y - D\theta_0\}\big(D - X'\pi_0\big)\big] = 0.$$

The first moment condition corresponds to the regression method, while the second corresponds to the so-called covariate balancing method. Importantly, the use of these non-orthogonal moment conditions generally does not produce an estimator for θ_0 that is $\sqrt{n}$-consistent and asymptotically normal uniformly in the class of approximately sparse models. This failure occurs because we are forced to use highly non-regular estimators to estimate the nuisance functions $X'\delta_0$ and $X'\pi_0$ in the $p \gg n$ setting. In fact, this failure would also occur with a low number of controls, including having only $p = 1$, whenever selection procedures that exclude irrelevant variables with very high probability are used to estimate the regression parameter δ_0 or the propensity score parameter π_0. For more discussion and documentation of this failure, see Leeb and Pötscher (2008a, 2008b), Pötscher (2009), and Belloni, Chernozhukov, and Hansen (2013, 2014). By contrast, constructing orthogonal moment conditions—involving the projection of both the outcome and the treatment onto the controls and thereby combining the regression and covariate balancing methods—makes it possible to achieve $\sqrt{n}$ consistency and asymptotic normality uniformly within a class of approximately sparse models.

EXAMPLE 2—Neyman Orthogonal Equations for a Class of Conditional Moment Problems: Next, consider the conditional moment restrictions framework studied by Chamberlain (1992):

$$\mathrm{E}\big[\varphi\big(W, \theta_0, g_0(X)\big)|X\big] = 0,$$

where X and W are random vectors with X being a sub-vector of W, $\theta \in \Theta \subset \mathbb{R}^d$ is a finite-dimensional parameter whose true value θ_0 is of interest, g is a functional nuisance parameter mapping the support of X into a convex set $V \subset \mathbb{R}^l$ whose true value is g_0, and φ is a known function with values in $\mathbb{R}^k$ for $k \geq d + l$.

Here we would like to build a score function $(\theta, h) \longmapsto \psi(W, \theta, h)$ for estimating θ_0, the true value of parameter θ, where h is a new nuisance parameter with true value h_0 that obeys the strong form of the orthogonality condition (5.5) and thus also its weak form (5.8). To this end, let $t \longmapsto \mathrm{E}[\varphi(W, \theta_0, t)|X]$ be a function mapping $\mathbb{R}^l$ into $\mathbb{R}^k$ and let $\gamma(X, \theta_0, g_0) = \partial_{t'}\mathrm{E}[\varphi(W, \theta_0, t)|X]|_{t=g_0(X)}$ be a $k \times l$ matrix of its derivatives. We will set $Z = X$ and $h(X) = \mathrm{vec}(g(X), \beta(X), \Sigma(X))$ where β is a function mapping the support of X into the space of $d \times k$ matrices, $\mathbb{R}^{d\times k}$, and Σ is the function mapping the support of X into the space of $k \times k$ matrices, $\mathbb{R}^{k\times k}$. Define the true value β_0 of β as

$$\beta_0(X) = A(X)\big(I - \Pi_0(X)\big),$$

where $A(X)$ is a $d \times k$ matrix of measurable transformations of X, I is the $k \times k$ identity matrix, and $\Pi_0(X) \neq I_{k\times k}$ is a $k \times k$ non-identity matrix with the property

$$\Pi_0(X)\Sigma_0^{-1/2}(X)\gamma(X, \theta_0, g_0) = \Sigma_0^{-1/2}(X)\gamma(X, \theta_0, g_0), \tag{5.14}$$

where Σ_0 is the true value of parameter Σ. For example, $\Pi_0(X)$ can be chosen to be an orthogonal projection matrix:

$$\begin{aligned}\Pi_0(X) = \big[&\Sigma_0(X)^{-1/2}\gamma(X, \theta_0, g_0)\\ &\times \big(\gamma(X, \theta_0, g_0)'\Sigma_0(X)^{-1}\gamma(X, \theta_0, g_0)\big)^{-1}\\ &\times \gamma(X, \theta_0, g_0)'\Sigma_0(X)^{-1/2}\big].\end{aligned}$$

Then an orthogonal score for the problem above can be constructed as

$$\begin{aligned}&\psi\big(W, \theta, h(X)\big) = \beta(X)\Sigma^{-1/2}(X)\varphi\big(Z, \theta, h(X)\big),\\ &h(X) = \mathrm{vec}\big(g(X), \beta(X), \Sigma(X)\big).\end{aligned}$$

It is straightforward to check that, under mild regularity conditions, the score function ψ satisfies $\mathrm{E}[\psi(W, \theta_0, h_0(X))] = 0$ for $h_0(X) = \mathrm{vec}(g_0(X), \beta_0(X), \Sigma_0(X))$ and also obeys the orthogonality condition

$$\partial_t \mathrm{E}_P\big[\psi(W, \theta_0, t)|X\big]\big|_{t=h_0(Z)} = 0, \quad \text{a.s.} \tag{5.15}$$

Furthermore, by setting

$$A(X) = \big(\partial_{\theta'} \mathrm{E}\big[\varphi\big(W, \theta, g_0(X)\big)|X\big]\big|_{\theta=\theta_0}\big)',$$

$$\Sigma_0(X) = \mathrm{E}\big[\varphi\big(W, \theta_0, g_0(X)\big)\varphi\big(W, \theta_0, g_0(X)\big)'|X\big],$$

and using $\Pi_0(X)$ suggested above, we obtain the efficient score ψ that yields an estimator of θ_0 achieving the semiparametric efficiency bound provided in Chamberlain (1992).

Here we would like to note that an analogous, though more involved, construction can be provided for the more general class of problems considered in Ai and Chen (2003) where the nuisance functions depend on the endogenous variables.

5.3. *Regularity Conditions and Results*

In what follows, we shall denote by δ, c_0, c, and C some positive constants. For a positive integer d, $[d]$ denotes the set $\{1, \dots, d\}$. We shall impose the following regularity conditions.

ASSUMPTION 5.1—Moment-Condition Problem: *Consider a random element W, taking values in a measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((W_{ui})_{u\in\mathcal{U}})_{i=1}^n$ consist of n i.i.d. copies of a random element $(W_u)_{u\in\mathcal{U}}$ which is generated as a suitably measurable transformation with respect to W and u. Uniformly for all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold*: (i) *The true parameter value θ_u obeys* (5.1) *and is interior relative to $\Theta_u \subset \Theta \subset \mathbb{R}^{d_\theta}$, namely there is a ball of radius δ centered at θ_u contained in Θ_u for all $u \in \mathcal{U}$, and Θ is compact*. (ii) *For $\nu := (\nu_k)_{k=1}^{d_\theta+d_t} = (\theta, t)$, each $j \in [d_\theta]$, and $u \in \mathcal{U}$, the map $\Theta_u \times T_u(Z_u) \ni \nu \longmapsto \mathrm{E}_P[\psi_{uj}(W_u, \nu)|Z_u]$ is twice continuously differentiable a.s. with derivatives obeying the integrability conditions specified in Assumption* 5.2. (iii) *For all $u \in \mathcal{U}$, the moment function ψ_u obeys the orthogonality condition given in Definition* 5.1 *for the set $\mathcal{H}_u = \mathcal{H}_{un}$ specified in Assumption* 5.3. (iv) *The following identifiability condition holds*: $\|\mathrm{E}_P[\psi_u(W_u, \theta, h_u(Z_u))]\| \geq 2^{-1}(\|J_u(\theta - \theta_u)\| \wedge c_0)$ *for all $\theta \in \Theta_u$, where the singular values of $J_u := \partial_\theta \mathrm{E}[\psi_u(W_u, \theta_u, h_u(Z_u))]$ lie between c and C for all $u \in \mathcal{U}$.*

The conditions of Assumption 5.1 are mild and standard in moment-condition problems. Assumption 5.1(iv) encodes sufficient global and local identifiability to obtain a rate result. The suitably measurable condition, defined in Appendix A, is a mild condition satisfied in most practical cases.

ASSUMPTION 5.2—Entropy and Smoothness: *The set $(\mathcal{U}, d_{\mathcal{U}})$ is a semimetric space such that $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leq C \log(\mathrm{e}/\epsilon) \vee 0$. Let $\alpha \in [1, 2]$, and let α_1 and α_2 be some positive constants. Uniformly for all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold*: (i) *The set of functions $\mathcal{F}_0 = \{\psi_{uj}(W_u, \theta_u, h_u(Z_u)) : j \in [d_\theta], u \in \mathcal{U}\}$, viewed as functions of W, is suitably measurable*; *has an envelope function $F_0(W) = \sup_{j\in[d_\theta], u\in\mathcal{U}, \nu\in\Theta_u\times T_u(Z_u)} |\psi_{uj}(W_u, \nu)|$ that is measurable with respect to W and obeys $\|F_0\|_{P,q} \leq C$, where $q \geq 4$ is a fixed constant*; *and has a uniform covering entropy obeying $\sup_Q \log N(\epsilon\|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leq C \log(\mathrm{e}/\epsilon) \vee 0$.*
(ii) *For all $j \in [d_\theta]$ and $k, r \in [d_\theta + d_t]$, and $\psi_{uj}(W) := \psi_{uj}(W_u, \theta_u, h_u(Z_u))$*:
(a) $\sup_{u\in\mathcal{U}, (\nu,\bar{\nu})\in(\Theta_u\times T_u(Z_u))^2} \mathrm{E}_P[(\psi_{uj}(W_u, \nu) - \psi_{uj}(W_u, \bar{\nu}))^2|Z_u]/\|\nu - \bar{\nu}\|^\alpha \leq C$, *$P$-a.s.*,
(b) $\sup_{d_{\mathcal{U}}(u,\bar{u})\leq\delta} \mathrm{E}_P[(\psi_{uj}(W) - \psi_{\bar{u}j}(W))^2] \leq C\delta^{\alpha_1}$, $\sup_{d_{\mathcal{U}}(u,\bar{u})\leq\delta} \|J_u - J_{\bar{u}}\| \leq C\delta^{\alpha_2}$,
(c) $\mathrm{E}_P \sup_{u\in\mathcal{U}, \nu\in\Theta_u\times T_u(Z_u)} |\partial_{\nu_r} \mathrm{E}_P[\psi_{uj}(W_u, \nu)|Z_u]|^2 \leq C$,
(d) $\sup_{u\in\mathcal{U}, \nu\in\Theta_u\times T_u(Z_u)} |\partial_{\nu_k}\partial_{\nu_r} \mathrm{E}_P[\psi_{uj}(W_u, \nu)|Z_u]| \leq C$, *$P$-a.s.*

Assumption 5.2 imposes smoothness and integrability conditions on various quantities derived from ψ_u. It also imposes conditions on the complexity of the relevant function classes.

In what follows, let $\Delta_n \searrow 0$, $\delta_n \searrow 0$, and $\tau_n \searrow 0$ be sequences of constants approaching zero from above at a speed at most polynomial in n (e.g., $\delta_n \geq 1/n^c$ for some $c > 0$).

ASSUMPTION 5.3—Estimation of Nuisance Functions: *The following conditions hold for each $n \geq n_0$ and all $P \in \mathcal{P}_n$. The estimated functions $\hat{h}_u = (\hat{h}_{um})_{m=1}^{d_t} \in \mathcal{H}_{un}$ with probability at least $1 - \Delta_n$, where $\mathcal{H}_{un}$ is the set of measurable maps $\mathcal{Z}_u \ni z \longmapsto h = (h_m)_{m=1}^{d_t}(z) \in T_u(z)$ such that*

$$\|h_m - h_{um}\|_{P,2} \leq \tau_n, \quad \tau_n^2 \sqrt{n} \leq \delta_n,$$

and whose complexity does not grow too quickly in the sense that $\mathcal{F}_1 = \{\psi_{uj}(W_u, \theta, h(Z_u)) : j \in [d_\theta], u \in \mathcal{U}, \theta \in \Theta_u, h \in \mathcal{H}_{un}\}$ is suitably measurable and its uniform covering entropy obeys

$$\sup_Q \log N\big(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}\big) \leq s_n\big(\log(a_n/\epsilon)\big) \vee 0,$$

where $F_1(W)$ is an envelope for $\mathcal{F}_1$ which is measurable with respect to W and satisfies $F_1(W) \leq F_0(W)$ for F_0 defined in Assumption 5.2. The complexity characteristics $a_n \geq$ $\max(n, \mathrm{e})$ *and $s_n \geq 1$ obey the growth conditions*

$$n^{-1/2}\big(\sqrt{s_n \log(a_n)} + n^{-1/2} s_n n^{1/q} \log(a_n)\big) \leq \tau_n \quad \textit{and}$$
$$\tau_n^{\alpha/2}\sqrt{s_n \log(a_n)} + s_n n^{1/q-1/2} \log(a_n) \log n \leq \delta_n,$$

where q and α are defined in Assumption 5.2.

COMMENT 5.4—On Rate and Entropy Rate Conditions: Assumption 5.3 imposes conditions on the estimation rate of the nuisance functions h_{um} and on the complexity of the functions sets that contain the estimators $\hat{h}_{um}$. This condition allows for a wide variety of modern modeling assumptions and regularization methods for function fitting, including both traditional methods and more recent statistical and machine learning methods. Within the approximately sparse framework, the index s_n corresponds to the maximum of the dimension of the approximating models and of the size of the selected models; and $a_n = p \vee n$. Under other frameworks, these parameters could be different; yet if they are well-behaved, then our results still apply. Thus, these results cover other frameworks, where structured assumptions other than approximate sparsity are used to make the estimation and modeling problem manageable. It is important to point out that the class $\mathcal{F}_1$ generally will not be Donsker because its entropy is allowed to increase with n. Allowing for non-Donsker classes is crucial for accommodating *modern*, *high-dimensional* estimation methods for the nuisance functions. This feature makes the conditions imposed here very different from the conditions imposed in various classical references on dealing with nonparametrically estimated nuisance functions; see, for example, van der Vaart and Wellner (1996), van der Vaart (1998), Kosorok (2008), and other references listed in the Introduction.

COMMENT 5.5—Removing Entropy Rate Conditions by Sample-Splitting: We can can set $s_n = 1$ and $a_n = e$ in Assumption 5.3 if we employ data-splitting. That is, under data-splitting the entropy condition becomes very weak, akin to that in parametric problems, facilitating the application of modern statistical and machine learning methods (e.g. random forest, boosted trees, deep neural nets, and their aggregated and hybrid versions) to estimate the nuisance functions. Thus, with data-splitting Assumption 5.3 only requires that the estimators of nuisance parameters attain sufficiently rapid rates of convergences τ_n, in particular $\tau_n = o(n^{-1/4})$ in smooth problems. Of course in practice we can not verify that these rates hold in a given problem, but the regularity conditions become more plausible with data-splitting than without it. Bickel (1982) employs the idea of data-splitting, namely setting aside a vanishing fraction of the sample to estimate the nuisance parameter, to set up adaptive estimators of the main parameter; see also van der Vaart (1998). This ensures that there is no asymptotic efficiency loss from data-splitting. Another method, which seems more practical, is to use the following cross-fitting approach: (1) split the sample into two equal parts, the auxiliary and main parts; (2) use the auxiliary part to estimate the nuisance parameter and the main part to estimate the target parameter, obtaining one estimator of the target parameter; (3) by reversing the roles of the main and auxiliary parts, obtain another estimator of the target parameter; and (4) average the two estimators of the target parameter to obtain the final estimator. Theorems 5.1 given below yields the properties of the final estimator. We refer to Chernozhukov et al. (2016) for further details, including the result that there is no asymptotic efficiency loss from data-splitting under cross-fitting.

The following theorem is one of the main results of the paper.

THEOREM 5.1—Uniform Functional Central Limit Theorem for a Continuum of Target Parameters in Moment-Condition Problems: *Under Assumptions* 5.1, 5.2, *and* 5.3, *for an estimator* $(\hat{\theta}_u)_{u\in\mathcal{U}}$ *that obeys equation* (5.4),

$$\sqrt{n}(\hat{\theta}_u - \theta_u)_{u\in\mathcal{U}} = (\mathbb{G}_n \bar{\psi}_u)_{u\in\mathcal{U}} + o_P(1)$$

in $\ell^\infty(\mathcal{U})^{d_\theta}$, *uniformly in* $P \in \mathcal{P}_n$, *where* $\bar{\psi}_u(W) := -J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u))$, *and*

$$Z_{n,P} := (\mathbb{G}_n \bar{\psi}_u)_{u\in\mathcal{U}} \rightsquigarrow Z_P := (\mathbb{G}_P \bar{\psi}_u)_{u\in\mathcal{U}} \quad \text{in} \quad \ell^\infty(\mathcal{U})^{d_\theta}, \text{ uniformly in } P \in \mathcal{P}_n,$$

where the paths of $u \longmapsto \mathbb{G}_P \bar{\psi}_u$ *are a.s. uniformly continuous on* $(\mathcal{U}, d_{\mathcal{U}})$ *and*

$$\sup_{P\in\mathcal{P}_n} \mathrm{E}_P \sup_{u\in\mathcal{U}} \|\mathbb{G}_P \bar{\psi}_u\| < \infty \quad \text{and}$$

$$\lim_{\delta\to 0} \sup_{P\in\mathcal{P}_n} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u,\bar{u})\le\delta} \|\mathbb{G}_P \bar{\psi}_u - \mathbb{G}_P \bar{\psi}_{\bar{u}}\| = 0.$$

COMMENT 5.6: It is important to mention here that this result on a continuum of parameters solving a continuum of moment conditions is completely new. The prior approaches dealing with continua of moment conditions with infinite-dimensional nuisance parameters, for example, the ones given in Chernozhukov and Hansen (2006) and Escanciano and Zhu (2013), impose Donsker conditions on the class of functions, following Andrews (1994a), that contain the values of the estimators of these nuisance functions. This approach is precluded in our setting because the resulting class of functions in our case has entropy that grows with the sample size and therefore is not Donsker. Hence, we develop a new approach to establishing the results which exploits the interplay between the rate of growth of entropy, the biases, and the size of the estimation error. In addition, the new approach allows for obtaining results that are uniform in P.

We can estimate the law of Z_P with the bootstrap law of

$$\hat{Z}^*_{n,P} := \sqrt{n}\big(\hat{\theta}^*_u - \hat{\theta}_u\big)_{u\in\mathcal{U}} := \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \hat{\psi}_u(W_i)\right)_{u\in\mathcal{U}}, \tag{5.16}$$

where $(\xi_i)_{i=1}^n$ are i.i.d. multipliers as defined in equation (3.20), $\hat{\psi}_u(W_i)$ is the estimated score

$$\hat{\psi}_u(W_i) := -\hat{J}_u^{-1}\psi_u\big(W_{ui}, \hat{\theta}_u, \hat{h}_u(Z_{ui})\big),$$

and $\hat{J}_u$ is a suitable estimator of J_u.[19] The bootstrap law is computed by drawing $(\xi_i)_{i=1}^n$ conditional on the data.

The following theorem shows that the multiplier bootstrap provides a valid approximation to the large sample law of $\sqrt{n}(\hat{\theta}_u - \theta_u)_{u\in\mathcal{U}}$.

THEOREM 5.2—Uniform Validity of Multiplier Bootstrap: *Suppose Assumptions* 5.1, 5.2, *and* 5.3 *hold*, *the estimator* $(\hat{\theta}_u)_{u\in\mathcal{U}}$ *obeys equation* (5.4), *and that the estimator* $(\hat{J}_u)_{u\in\mathcal{U}}$ *obeys the following condition*: *uniformly in* $P \in \mathcal{P}_n$ *with probability* $1-\delta_n$, $\sup_{u\in\mathcal{U}}\|\hat{J}_u - J_u\| \le \Delta_n$. *Then*,

$$\hat{Z}^*_{n,P} \rightsquigarrow_B Z_P \quad \textit{in} \quad \ell^\infty(\mathcal{U})^{d_\theta}, \textit{ uniformly in } P \in \mathcal{P}_n.$$

We next derive the large sample distribution and validity of the multiplier bootstrap for the estimator $\hat{\Delta} := \phi(\hat{\theta}) := \phi((\hat{\theta}_u)_{u\in\mathcal{U}})$ of the functional $\Delta := \phi(\theta^0) = \phi((\theta_u)_{u\in\mathcal{U}})$ using the functional delta method. The functional $\theta^0 \longmapsto \phi(\theta^0)$ is defined as a uniformly Hadamard differentiable transform of $\theta^0 = (\theta_u)_{u\in\mathcal{U}}$. The following result gives the large sample law of $\sqrt{n}(\hat{\Delta} - \Delta)$, the properly normalized estimator. It also shows that the bootstrap law of $\sqrt{n}(\hat{\Delta}^* - \hat{\Delta})$, computed conditionally on the data, is consistent for the large sample law of $\sqrt{n}(\hat{\Delta} - \Delta)$. Here $\hat{\Delta}^* := \phi(\hat{\theta}^*) = \phi((\hat{\theta}^*)_{u\in\mathcal{U}})$ is the bootstrap version of $\hat{\Delta}$, and $\hat{\theta}^*_u = \hat{\theta}_u + n^{-1}\sum_{i=1}^n \xi_i\hat{\psi}_u(W_i)$ is the multiplier bootstrap version of $\hat{\theta}_u$ defined via equation (5.16).

THEOREM 5.3—Uniform Limit Theory and Validity of Multiplier Bootstrap for Smooth Functionals of θ: *Suppose that for each* $P \in \mathcal{P} := \bigcup_{n\ge n_0}\mathcal{P}_n$, $\theta^0 = \theta^0_P$ *is an element of a compact set* $\mathbb{D}_\theta$. *Suppose* $\theta \longmapsto \phi(\theta)$, *a functional of interest mapping* $\mathbb{D}_\phi \subset \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}$ *to* $\ell^\infty(\mathcal{Q})$, *where* $\mathbb{D}_\theta \subset \mathbb{D}_\phi$, *is Hadamard differentiable in* θ *tangentially to* $\mathbb{D}_0 = UC(\mathcal{U})^{d_\theta}$ *uniformly in* $\theta \in \mathbb{D}_\theta$, *with the linear derivative map* $\phi'_\theta : \mathbb{D}_0 \longmapsto \mathbb{D}$ *such that the mapping* $(\theta, h) \longmapsto \phi'_\theta(h)$ *from* $\mathbb{D}_\theta \times \mathbb{D}_0$ *to* $\ell^\infty(\mathcal{Q})$ *is continuous*. *Then*,

$$\sqrt{n}(\hat{\Delta} - \Delta) \rightsquigarrow T_P := \phi'_{\theta^0_P}(Z_P) \quad \textit{in} \quad \ell^\infty(\mathcal{Q}), \textit{ uniformly in } P \in \mathcal{P}_n, \tag{5.17}$$

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

$$\sqrt{n}\big(\hat{\Delta}^* - \hat{\Delta}\big) \rightsquigarrow_B T_P \quad \textit{in} \quad \ell^\infty(\mathcal{Q}), \textit{ uniformly in } P \in \mathcal{P}_n. \tag{5.18}$$

[19]We do not discuss the estimation of J_u since it is often a problem-specific matter. In Section 3, J_u was equal to minus the identity matrix, so we did not need to estimate it.

To derive Theorem 5.3, we strengthen the usual notion of Hadamard differentiability to a uniform notion introduced in Definition B.1. Theorems B.3 and B.4 show that this uniform Hadamard differentiability is sufficient to guarantee the validity of the functional delta uniformly in P. These new uniform functional delta method theorems may be of independent interest.

6. THEORY: LASSO AND POST-LASSO FOR FUNCTIONAL RESPONSE DATA

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

6.1. *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 $u\in\mathcal{U}$,

$$\mathrm{E}_P[Y_u|X] = \Lambda\big(f(X)'\theta_u\big) + r_u(X), \tag{6.1}$$

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 known link function. The notation in this section differs from the rest of the paper with Y_u and X denoting a generic response and a generic vector of covariates to facilitate the application of these results to other contexts. We only consider the linear link function, $\Lambda(t)=t$, and the logistic link function, $\Lambda(t)=\exp(t)/\{1+\exp(t)\}$, in detail.

Considering the logistic link is useful when the functional response is binary, though the linear link can be used in that case as well under some conditions. For example, it is useful for estimating a high-dimensional generalization of the distributional regression models considered in Chernozhukov, Fernández-Val, and Melly (2013) where the response variable is the continuum $(Y_u = 1(Y\le u))_{u\in\mathcal{U}}$. Even though we focus on these two cases, we note that the principles discussed here apply to many other M-estimators with convex (or approximately convex) criterion functions. In the remainder of the section, we discuss and establish results for ℓ_1-penalized and post-model selection estimators of $(\theta_u)_{u\in\mathcal{U}}$ 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 we have n i.i.d. observations from d.g.p.'s where (6.1) holds, $\{(Y_{ui})_{u\in\mathcal{U}}, X_i)\}_{i=1}^n$, available for estimating $(\theta_u)_{u\in\mathcal{U}}$. For each $u\in\mathcal{U}$, penalty level λ, and diagonal matrix of penalty loadings $\hat\Psi_u$, we define the Lasso estimator as

$$\hat\theta_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_n\big[M\big(Y_u, f(X)'\theta\big)\big] + \frac{\lambda}{n}\|\hat\Psi_u\theta\|_1, \tag{6.2}$$

where $M(y,t)=\frac{1}{2}(y-\Lambda(t))^2$ 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 the logistic link function for binary response data. For each $u\in\mathcal{U}$, the Post-Lasso estimator based on a set of covariates $\tilde T_u$ is then defined as

$$\tilde\theta_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_n\big[M\big(Y_u, f(X)'\theta\big)\big] : \mathrm{supp}(\theta)\subseteq \tilde T_u, \tag{6.3}$$

where the set $\tilde{T}_u$ contains $\text{supp}(\hat{\theta}_u)$ and may also contain additional variables deemed as important.[20] We will set $\tilde{T}_u = \text{supp}(\hat{\theta}_u)$ unless otherwise noted.

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

$$\frac{\lambda}{n} \geq c \sup_{u\in\mathcal{U}} \left\| \hat{\Psi}_u^{-1} \mathbb{E}_n\left[\partial_\theta M\left(Y_u, f(X)'\theta_u\right)\right]\right\|_\infty, \tag{6.4}$$

where $c > 1$ is a fixed constant. When $\mathcal{U}$ is a singleton, the strategy above is similar to Bickel, Ritov, and Tsybakov (2009), Belloni and Chernozhukov (2013), and Belloni, Chernozhukov, and Wang (2011), who used an analog of (6.4) to derive the properties of Lasso and Post-Lasso. When $\mathcal{U}$ is not a singleton, this strategy was first employed in the context of ℓ_1-penalized quantile regression processes by Belloni and Chernozhukov (2011).

To implement (6.4), we propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}\left(1 - \gamma/\left\{2pn^{d_u}\right\}\right), \tag{6.5}$$

where d_u is the dimension of $\mathcal{U}$, $1 - \gamma$ with $\gamma = o(1)$ is a confidence level associated with the probability of event (6.4), and $c > 1$ is a slack constant.[21] When implementing the estimators, we set $c = 1.1$ and $\gamma = 0.1/\log(n)$, which is theoretically motivated and practically tested in an extensive set of simulation experiments in Belloni, Chernozhukov, and Hansen (2014). In addition to the penalty parameter λ, we also need to construct a penalty loading matrix $\hat{\Psi}_u = \text{diag}(\{\hat{l}_{uj}, j = 1, \ldots, p\})$. This loading matrix can be formed according to the following iterative algorithm.

ALGORITHM 6.1—Estimation of Penalty Loadings: *Choose* $\gamma \in [1/n, \min\{1/\log n, pn^{d_u-1}\}]$ *and* $c > 1$ *to form* λ *as defined in* (6.5), *and choose a constant* $K \geq 1$ *as an upper bound on the number of iterations*. (0) *Set* $k = 0$, *and initialize* $\hat{l}_{uj,0}$ *for each* $j = 1, \ldots, p$. *For the linear link function, set* $\hat{l}_{uj,0} = \{\mathbb{E}_n[f_j^2(X)(Y_u - \bar{Y}_u)^2]\}^{1/2}$ *with* $\bar{Y}_u = \mathbb{E}_n[Y_u]$. *For the logistic link function, set* $\hat{l}_{uj,0} = \frac{1}{2}\{\mathbb{E}_n[f_j^2(X)]\}^{1/2}$. (1) *Compute the Lasso and Post-Lasso estimators*, $\hat{\theta}_u$ *and* $\tilde{\theta}_u$, *based on* $\hat{\Psi}_u = \text{diag}(\{\hat{l}_{uj,k}, j = 1, \ldots, p\})$. (2) *Set* $\hat{l}_{uj,k+1} := \{\mathbb{E}_n[f_j^2(X)(Y_u - \Lambda(f(X)'\tilde{\theta}_u))^2]\}^{1/2}$. (3) *If* $k > K$, *stop*; *otherwise set* $k \leftarrow k + 1$ *and go to step* (1).

6.2. *Properties of a Continuum of Lasso and Post-Lasso: Linear Link*

We provide sufficient conditions for establishing good performance of the estimators discussed above when the linear link function is used. In the statement of the following assumption, $\delta_n \searrow 0$ and $\Delta_n \searrow 0$ are fixed sequences approaching zero from above at a speed at most polynomial in n (e.g., $\delta_n \geq 1/n^c$ for some $c > 0$), $\ell_n := \log n$, and c, C, κ', κ'' and $\nu \in (0, 1]$ are positive finite constants.

[20]The total number of additional variables $\hat{s}_a$ should also obey the same growth conditions that s obeys. For example, if the additional variables are chosen so that $\hat{s}_a \lesssim \|\hat{\theta}_u\|_0$, the growth condition is satisfied with probability going to 1 for the designs covered by Assumptions 6.1 and 6.2. See also Belloni, Chernozhukov, and Hansen (2014) for a discussion on choosing additional variables.

[21]When the set $\mathcal{U}$ is a singleton, one can use the penalty level in (6.5) with $d_u = 0$. This choice corresponds to that used in Belloni, Chernozhukov, and Hansen (2014).

ASSUMPTION 6.1: *Consider a random element W taking values in a measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((Y_{ui})_{u\in\mathcal{U}}, X_i)_{i=1}^n$ consist of n i.i.d. copies of random element $((Y_u)_{u\in\mathcal{U}}, X)$, which is generated as a suitably measurable transformation of W and u. The model* (6.1) *holds with linear link $t \longmapsto \Lambda(t) = t$ for all $u \in \mathcal{U} \subset [0,1]^{d_u}$, where d_u is fixed and $\mathcal{U}$ is equipped with the semimetric $d_{\mathcal{U}}$. Uniformly for all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold.* (i) *The model* (6.1) *is approximately sparse with sparsity index obeying* $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \leq s$ *and the growth restriction* $\log(p \vee n) \leq \delta_n n^{1/3}$. (ii) *The set $\mathcal{U}$ has uniform covering entropy obeying* $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leq d_u \log(1/\epsilon) \vee 0$, *and the collection $(\zeta_u = Y_u - \mathrm{E}_P[Y_u|X], r_u)_{u\in\mathcal{U}}$ are suitably measurable transformations of W and u.* (iii) *Uniformly over $u \in \mathcal{U}$, the moments of the model are boundedly heteroscedastic, namely* $c \leq \mathrm{E}_P[\zeta_u^2|X] \leq C$ *a.s., and* $\max_{j\leq p}\mathrm{E}_P[|f_j(X)\zeta_u|^3 + |f_j(X)Y_u|^3] \leq C$. (iv) *For a fixed $\nu > 0$ and a sequence K_n, the dictionary functions, approximation errors, and empirical errors obey the following regularity conditions*: (a) $c \leq \mathrm{E}_P[f_j^2(X)] \leq C$, $j = 1, \dots, p$; $\max_{j\leq p} |f_j(X)| \leq K_n$ *a.s.*; $K_n^2 s \log(p \vee n) \leq \delta_n n$. (b) *With probability* $1 - \Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \leq Cs\log(p \vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\leq p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \vee |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)Y_u^2]| \leq \delta_n$; $\log^{1/2}(p \vee n)\sup_{d_{\mathcal{U}}(u,u')\leq 1/n} \max_{j\leq p}\{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2} \leq \delta_n$, *and* $\sup_{d_{\mathcal{U}}(u,u')\leq 1/n}\|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \leq \delta_n n^{-1/2}$. (c) *With probability* $1 - \Delta_n$, *the empirical minimum and maximum sparse eigenvalues are bounded from zero and above, namely* $\kappa' \leq \inf_{\|\delta\|_0\leq s\ell_n, \|\delta\|=1} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leq \sup_{\|\delta\|_0\leq s\ell_n, \|\delta\|=1} \|f(X)'\delta\|_{\mathbb{P}_n,2} \leq \kappa''$.

Assumption 6.1 is only a set of sufficient conditions. The finite sample results in the Supplemental Material allow for more general conditions (e.g., d_u can grow with the sample size). We verify that the more technical conditions in Assumption 6.1(iv)(b) hold in a variety of cases; see Lemma J.2 in Appendix J in the Supplemental Material. Under Assumption 6.1, we establish results on the performance of the estimators (6.2) and (6.3) for the linear link function case that hold uniformly over $u \in \mathcal{U}$ and $P \in \mathcal{P}_n$.

THEOREM 6.1—Rates and Sparsity for Functional Responses Under Linear Link: *Under Assumption* 6.1 *and setting the penalty and loadings as in Algorithm* 6.1, *for all n large enough, uniformly for all $P \in \mathcal{P}_n$ with P_P probability $1 - o(1)$, for some constant $\bar{C}$, the Lasso estimator $\hat{\theta}_u$ is uniformly sparse,* $\sup_{u\in\mathcal{U}} \|\hat{\theta}_u\|_0 \leq \bar{C}s$, *and the following performance bounds hold*:

$$\sup_{u\in\mathcal{U}}\big\|f(X)'(\hat{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \leq \bar{C}\sqrt{\frac{s\log(p \vee n)}{n}} \quad \textit{and}$$

$$\sup_{u\in\mathcal{U}} \|\hat{\theta}_u - \theta_u\|_1 \leq \bar{C}\sqrt{\frac{s^2\log(p \vee n)}{n}}.$$

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

$$\sup_{u\in\mathcal{U}}\big\|f(X)'(\tilde{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \leq \bar{C}\sqrt{\frac{s\log(p \vee n)}{n}} \quad \textit{and}$$

$$\sup_{u\in\mathcal{U}} \|\tilde{\theta}_u - \theta_u\|_1 \leq \bar{C}\sqrt{\frac{s^2\log(p \vee n)}{n}}.$$

We note that the performance bounds are exactly of the type used in Assumption 4.2 (see also Assumption I.1 in the Supplemental Material). Indeed, under the condition $s^2 \log^2(p \vee n) \log^2 n \le \delta_n n$, the rate of convergence established in Theorem 6.1 yields $\sqrt{s \log(p \vee n)/n} \le o(n^{-1/4})$.

6.3. *Properties of Lasso and Post-Lasso Estimators: Logistic Link*

We provide sufficient conditions to state results on the performance of the estimators discussed above for the logistic link function. Consider the fixed sequences $\delta_n \searrow 0$ and $\Delta_n \searrow 0$ approaching zero from above at a speed at most polynomial in n, $\ell_n := \log n$, and the positive finite constants c, C, κ', κ'', and $\underline{c} \le 1/2$.

ASSUMPTION 6.2: *Consider a random element W taking values in a measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with law determined by a probability measure $P \in \mathcal{P}_n$. The observed data $((Y_{ui})_{u\in\mathcal{U}}, X_i)_{i=1}^n$ consist of n i.i.d. copies of random element $((Y_u)_{u\in\mathcal{U}}, X)$, which is generated as a suitably measurable transformation of W and u. The model (6.1) holds with $Y_{ui} \in \{0, 1\}$ with the logistic link $t \longmapsto \Lambda(t) = \exp(t)/\{1 + \exp(t)\}$ for each $u \in \mathcal{U} \subset [0, 1]^{d_u}$, where d_u is fixed and $\mathcal{U}$ is equipped with the semimetric $d_{\mathcal{U}}$. Uniformly for all $n \ge n_0$ and $P \in \mathcal{P}_n$, the following conditions hold.* (i) *The model (6.1) is approximately sparse with sparsity index obeying* $\sup_{u\in\mathcal{U}} \|\theta_u\|_0 \le s$ *and the growth restriction* $\log(p \vee n) \le \delta_n n^{1/3}$. (ii) *The set $\mathcal{U}$ has uniform covering entropy obeying* $\log N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \le d_u \log(1/\epsilon) \vee 0$, *and the collection $(\zeta_u = Y_u - \mathrm{E}_P[Y_u|X], r_u)_{u\in\mathcal{U}}$ is a suitably measurable transformation of W and u.* (iii) *Uniformly over $u \in \mathcal{U}$, the moments of the model satisfy* $\max_{j\le p} \mathrm{E}_P[|f_j(X)|^3] \le C$, *and* $\underline{c} \le \mathrm{E}_P[Y_u|X] \le 1 - \underline{c}$ *a.s.* (iv) *For a sequence K_n, the dictionary functions, approximation errors, and empirical errors obey the following boundedness and empirical regularity conditions:* (a) $\sup_{u\in\mathcal{U}} |r_u(X)| \le \delta_n$ *a.s.*; $c \le \mathrm{E}_P[f_j^2(X)] \le C$, $j = 1, \dots, p$; $\max_{j\le p} |f_j(X)| \le K_n$ *a.s.*; *and* $K_n^2 s^2 \log^2(p \vee n) \le \delta_n n$. (b) *With probability* $1 - \Delta_n$, $\sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2(X)] \le C s \log(p \vee n)/n$; $\sup_{u\in\mathcal{U}} \max_{j\le p} |(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]| \le \delta_n$; $\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \{\mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2]\}^{1/2} \le \delta_n$, *and* $\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\le 1/n} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \le \delta_n n^{-1/2}$. (c) *With probability* $1 - \Delta_n$, *the empirical minimum and maximum sparse eigenvalues are bounded from zero and above*: $\kappa' \le \inf_{\|\delta\|_0 \le s\ell_n, \|\delta\|=1} \|f(X)'\delta\|_{\mathbb{P}_n,2} \le \sup_{\|\delta\|_0 \le s\ell_n, \|\delta\|=1} \|f(X)'\delta\|_{\mathbb{P}_n,2} \le \kappa''$.

The following result characterizes the performance of the estimators (6.2) and (6.3) for the logistic link function case under Assumption 6.2.

THEOREM 6.2—Rates and Sparsity for Functional Response Under Logistic Link: *Under Assumption 6.2 and setting the penalty and loadings as in Algorithm 6.1, 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 $\bar{C}$:*

$$\sup_{u\in\mathcal{U}} \big\| f(X)'(\hat{\theta}_u - \theta_u) \big\|_{\mathbb{P}_n,2} \le \bar{C} \sqrt{\frac{s \log(p \vee n)}{n}} \quad \textit{and}$$

$$\sup_{u\in\mathcal{U}} \|\hat{\theta}_u - \theta_u\|_1 \le \bar{C} \sqrt{\frac{s^2 \log(p \vee n)}{n}},$$

and the estimator is uniformly sparse: $\sup_{u\in\mathcal{U}} \|\hat{\theta}_u\|_0 \le \bar{C}s$. *For all n large enough, uniformly for all $P \in \mathcal{P}_n$, with P_P probability $1 - o(1)$, the Post-Lasso estimator corresponding to $\hat{\theta}_u$*

obeys

$$\sup_{u\in\mathcal{U}}\|f(X)'(\tilde{\theta}_u-\theta_u)\|_{\mathbb{P}_n,2}\leq\bar{C}\sqrt{\frac{s\log(p\vee n)}{n}}\quad \textit{and}$$

$$\sup_{u\in\mathcal{U}}\|\tilde{\theta}_u-\theta_u\|_1\leq\bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

COMMENT 6.1: The performance bounds derived in Theorem 6.2 satisfy the conditions of Assumption 4.2 (see also Assumption I.1 in the Supplemental Material). Moreover, since the link function is 1-Lipschitz in the logistic case and the approximation errors are assumed to be small, the results above establish the same rates of convergence for estimators of the conditional probabilities; for example,

$$\sup_{u\in\mathcal{U}}\|\mathrm{E}_P[Y_u|X]-\Lambda\big(f(X)'\hat{\theta}_u\big)\|_{\mathbb{P}_n,2}\leq\bar{C}\sqrt{\frac{s\log(p\vee n)}{n}}.$$

7. APPLICATION: THE EFFECT OF 401(K) PARTICIPATION ON ASSET HOLDINGS

As a practical illustration of the methods developed in this paper, we consider estimation of the effect of 401(k) eligibility and participation on accumulated assets as in Abadie (2003) and Chernozhukov and Hansen (2004). Our goal here is to illustrate the estimation results and inference statements and to make the following points that underscore our theoretical findings: (1) In a *low-dimensional setting*, where the number of controls is low and therefore there is no need for selection, our robust post-selection inference methods perform well. That is, the results of our methods agree with the results of standard methods that do not employ any selection. (2) In a *high-dimensional* setting, where there are (moderately) many controls, our post-selection inference methods perform well, producing well-behaved estimates and confidence intervals compared to the erratic estimates and confidence intervals produced by standard methods that do not employ selection as a means of regularization. (3) Finally, in a *very-high-dimensional* setting, where the number of controls is comparable to the sample size, the standard methods break down completely, while our methods still produce well-behaved estimates and confidence intervals. These findings are in line with our theoretical results about uniform validity of our inference methods.

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 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 high 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) adopted the strategy detailed in Poterba, Venti, and Wise (1994, 1995, 1996, 2001) and Benjamin (2003), who used data from the 1991 Survey of Income and Program Participation, and argued that eligibility for enrolling in a 401(k) plan in these 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) eligibility could be directly estimated by appropriate comparison across eligible and ineligible individuals.[22] Abadie (2003), Chernozhukov and Hansen (2004), and Ogburn, Rotnitzky, and Robins (2015) used this argument for the exogeneity of eligibility conditional on controls to argue that 401(k) eligibility provides a valid instrument for 401(k) participation and employed 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 financial wealth using high-dimensional methods. A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility may only be taken as exogenous after conditioning on income. Both Abadie (2003) and Chernozhukov and Hansen (2004) adopted this argument but controlled 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, the 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 previous research which implicitly assumes 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 Chernozhukov and Hansen (2004). The data consist of 9915 observations at the household level drawn from the 1991 SIPP. We use net financial assets as the outcome variable, Y, in our analysis. Our treatment variable, D, is an indicator for having positive 401(k) balances; and our instrument, Z, is an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, 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 can be found in Chernozhukov and Hansen (2004).

We present detailed results for three different sets of controls $f(X)$. The first specification uses indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, second-order polynomials in family size and education, a third-order polynomial in age, and a quadratic spline in income with six break points[23] (Quadratic Spline specification). The second specification augments the Quadratic Spline specification by interacting all the non-income variables with each term in the income spline (Quadratic Spline Plus Interactions specification). The final specification forms a larger set of potential controls by starting with all of the variables from the Quadratic Spline specification and forming all two-way interactions between all of the

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

[23]Specifically, we allow for income, income-squared, and then interact these two variables with seven dummies for the categories formed by the cut points.

non-income variables. The set of main effects and interactions of all non-income variables is then fully interacted with all of the income terms (Quadratic Spline Plus Many Interactions specification).[24] The dimensions of the set of controls are thus 35, 311, and 1756 for the Quadratic Spline, Quadratic Spline Plus Interactions, and Quadratic Spline Plus Many Interactions specification, respectively. For methods that do not use variable selection, we use 32, 272, and 1526 variables resulting from removing terms that are perfectly collinear. We refer to the specification without interactions as *low-* p, to the specification with only income interactions as *high-* p, and to the specification with all two-way interactions further interacted with income as *very-high-* p.

We report a variety of results for each specification. Under the maintained assumption that 401(k) eligibility may be taken as exogenous after controlling for the variables defined in the preceding paragraph, we can use the methods of this paper to estimate intention to treat effects of 401(k) eligibility by setting 401(k) eligibility as $D = Z$. We report the estimated average intention to treat and average intention to treat on the treated as the ATE and ATE-T, and we report estimates of quantile intention to treat and quantile intention to treat on the treated effects as QTE and QTE-T. We also directly apply the results of this paper to estimate effects of 401(k) participation, reporting estimates of the LATE, LATE-T, LQTE, and LQTE-T for each specification.[25] For comparison, we also report estimates of the eligibility effect from the linear model without selection and with selection using the approach of Belloni, Chernozhukov, and Hansen (2014) and estimates of the participation effect from linear instrumental variables estimation without selection and with selection as in Chernozhukov, Hansen, and Spindler (2015a).

Estimation of all these treatment effects depends on first-stage estimates of reduced-form functions as detailed in Section 3. We estimate reduced-form functions where the outcome is continuous using ordinary least squares when no model selection is used or Post-Lasso when selection is used. We estimate reduced-form functions where the outcome is binary by logistic regression when no model selection is used or Post-ℓ_1-penalized logistic regression when selection is used. We only report selection-based estimates in the very-high-p setting.[26] We refer to Appendix F for detailed discussion of implementing our approach in this example.

Estimates of the ATE, ATE-T, LATE, and LATE-T as well as the coefficient on 401(k) eligibility from the linear model and coefficient on 401(k) participation in the linear IV model are given in Table I. In this table, we provide point estimates for each of the three sets of controls with and without variable selection. We report conventional heteroscedasticity consistent standard error estimates for the linear model and linear IV coefficient. For the ATE, ATE-T, LATE, and LATE-T, we report both analytic and multiplier bootstrap standard errors. The bootstrap standard errors are based on 500 bootstrap replications with Mammen (1993) weights as multipliers.

[24]The specifications are motivated by the original specification used in Abadie (2003), Benjamin (2003), and Chernozhukov and Hansen (2004) allowing for data-dependent accommodation of nonlinearity. We report results based on the exact specification used in previous papers in the Supplemental Material.

[25]We note that because of one-sided compliance, the local effects for the treated actually coincide with population effects for the treated; see Frölich and Melly (2013).

[26]The estimated propensity score shows up in the denominator of the efficient moment conditions. As is conventional, we use trimming to keep the denominator bounded away from zero with trimming set to 10^{-12}. Trimming occurs in the Quadratic Spline Plus Interactions (12 observations trimmed) and Quadratic Spline Plus Many Interactions specifications (9915 observations trimmed) when selection is not done. Trimming never occurs in the selection-based estimates in this example. We choose not to report unregularized estimates in the very-high-p specification since all observations are trimmed and, in fact, have estimated propensity scores of either 0 or 1.

TABLE I

ESTIMATES AND STANDARD ERRORS OF AVERAGE EFFECTS[a]

Specification			Exogenous: 401(k) Eligibility			Endogenous: 401(k) Participation		
Series Approximation	Dimension	Selection	Linear Model	ATE	ATE-T	Linear IV	LATE	LATE-T
Quadratic Spline	35 (32)	N	8997	8093	11,250	12,926	11,579	15,969
			(1252)	(1082)	(1513)	(1796)	(1548)	(2148)
				{967}	{1423}		{1413}	{2194}
Quadratic Spline	35 (32)	Y	8967	7614	10,257	12,890	10,937	14,560
			(1270)	(1224)	(1776)	(1821)	(1758)	(2520)
				{1234}	{1676}		{1709}	{2576}
Quadratic Spline Plus Interactions	311 (272)	N	9019	11,775	11,740	12,973	17,529	16,664
			(1258)	(4202)	(1779)	(1804)	(6256)	(2526)
				{4202}	{1757}		{6249}	{2558}
Quadratic Spline Plus Interactions	311 (272)	Y	8307	7077	8830	11,784	10,168	12,533
			(1313)	(1358)	(2133)	(1995)	(1952)	(3027)
				{1237}	{2105}		{1963}	{2818}
Quadratic Spline Plus Many Interactions	1756 (1526)	N	8860	–	–	12,827	–	–
			(1358)	–	–	(1960)	–	–
				–	–		–	–
Quadratic Spline Plus Many Interactions	1756 (1526)	Y	8536	7848	9602	10,671	11,267	13,629
			(1321)	(1317)	(2047)	(2001)	(1890)	(2906)
				{1334}	{1894}		{1835}	{2862}

[a]The sample is drawn from the 1991 SIPP and consists of 9915 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. Quadratic Spline uses indicators for marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status; a third-order polynomial in age; second-order polynomials in education and family size; and a piecewise quadratic polynomial in income with six break points. The "Quadratic Spline Plus Interactions" specification include all first-order interactions between the income variables and the remaining variables. The specification denoted "Quadratic Spline Plus Many Interactions" takes all first-order interactions between all non-income variables and then fully interacts these interactions as well as the main effects with all income variables. Analytic standard errors are given in parentheses. Bootstrap standard errors based on 500 repetitions with Mammen (1993) multipliers are given in braces.

Looking first at the two sets of standard error estimates for the average treatment effect estimates, we see that the bootstrap and analytic standard errors are quite similar and that one would not draw substantively different conclusions from using one versus the other. We also see that estimates of the effect of 401(k) eligibility using the linear model and estimates of the effect of 401(k) participation using the linear IV model are broadly consistent with each other across all specifications and regardless of whether or not variable selection is done. We also have that the estimates of the ATE, ATE-T, LATE, and LATE-T are very similar regardless of whether selection is used in the low-p Quadratic Spline specification. The ATE and ATE-T both indicate a positive and significant average effect of 401(k) eligibility; and the LATE and LATE-T suggest positive and significant effects of 401(k) participation for compliers. The similarity in the low-p case is reassuring as it illustrates that there is little impact of variable selection relative to simply including everything in a low-dimensional setting.[27]

We observe somewhat different results in the Quadratic Spline Plus Interactions specification. For both the ATE and the LATE in the Quadratic Spline Plus Interactions case, we see a substantially larger point estimate without selection than with selection, with the selection results being similar to those obtained in the low-p case. Along with the larger point estimate, we also see that the estimated standard errors in the no-selection case for the ATE and LATE are roughly three times larger than the standard errors in the selection case. For the ATE-T and LATE-T in the Quadratic Spline Plus Interactions case, point estimates following selection are notably smaller than without selection but estimated standard errors after selection are somewhat larger. We note that one might suspect estimated standard errors for all of the estimators without selection to be substantially downward biased in this case due to the use of many control variables without regularization as in Cattaneo, Jansson, and Newey (2010). Finally, we see a large difference in the Orthogonal Polynomials Plus Many Interactions specifications as estimates cannot even be computed reliably without selection due to severe overfitting: The estimated propensity score is either 0 or 1 for every observation.

We provide estimates of the QTE and QTE-T in Figure 1 and estimates of the LQTE and LQTE-T in Figure 2. The left column of Figure 1 gives results for the QTE, and the right column displays the results for the QTE-T. Similarly, the left and right columns of Figure 2 provide the LQTE and LQTE-T, respectively. We give the results for the Quadratic Spline, Quadratic Spline Plus Interactions, and Quadratic Spline Plus Many Interactions specification in the top row, middle row, and bottom row, respectively. In each graphic, we use solid lines for point estimates and report uniform 95% confidence intervals with dashed lines.

Looking across the figures, we see a similar pattern to that seen for the estimates of the average effects in that the selection-based estimates are stable across all specifications and are very similar to the estimates obtained without selection from the baseline low-p Quadratic Spline specification. In the more flexible Quadratic Spline plus Interactions specification, the estimates that do not make use of selection behave somewhat erratically. This erratic behavior is especially apparent in the estimated LQTE of 401(k) participation where we observe that small changes in the quantile index may result in large swings in

[27]In the low-dimensional setting, using all available controls is semiparametrically efficient and allows uniformly valid inference. Thus, the similarity between the results in this case is an important feature of our method which results from our reliance on low-bias moment functions and sensible variable selection devices to produce semiparametrically efficient estimators and uniformly valid inference statements *following* model selection.

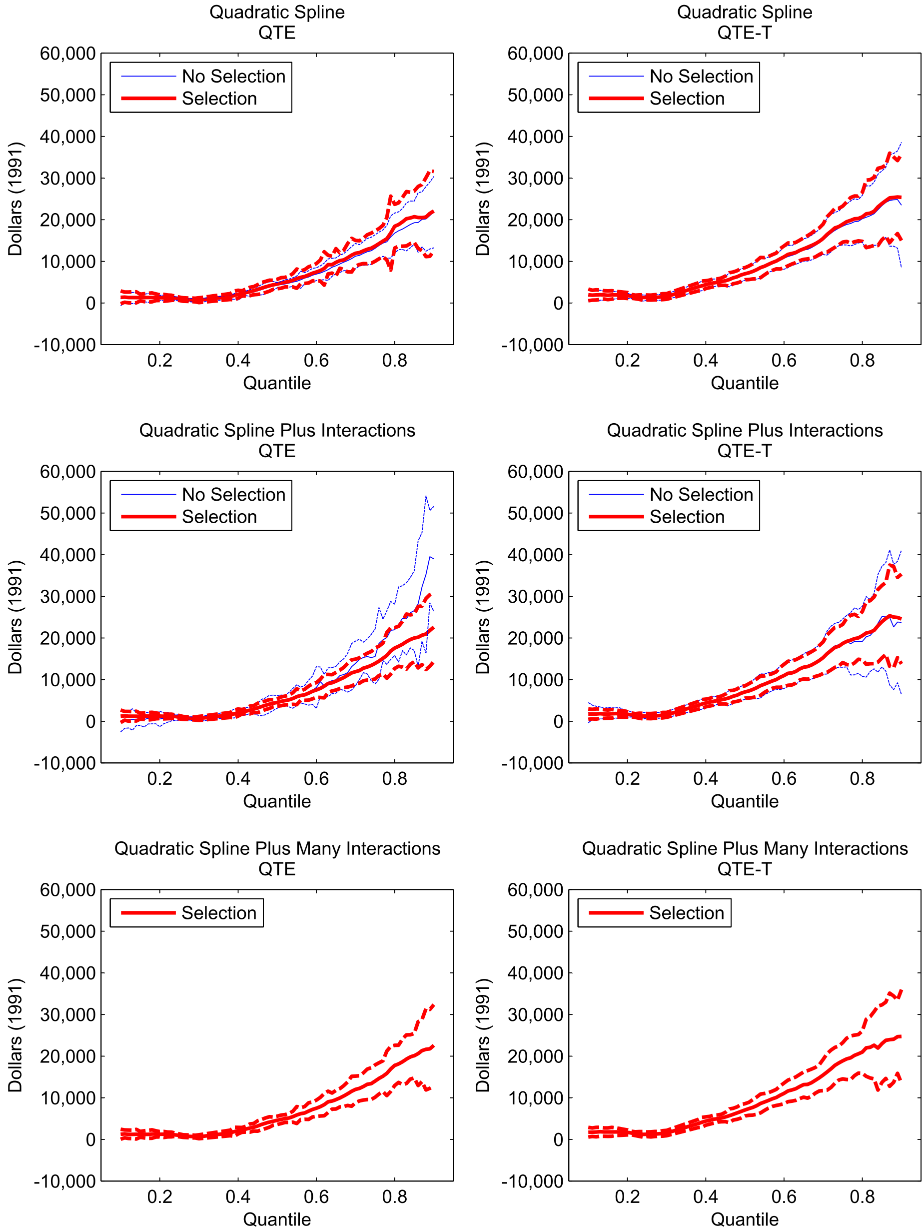

FIGURE 1.—QTE and QTE-T estimates of the effect of 401(k) eligibility on net financial assets.

the point estimate of the LQTE and estimated standard errors are quite large. Again, this erratic behavior is likely due to overfitting due to the large set of variables considered. As with the average effects, estimated quantile effects without selection in the Quadratic Spline Plus Many Interactions specification are not reported as the estimated propensity score is always 0 or 1.

If we focus on the LQTE and LQTE-T estimated from variable selection methods, we find that 401(k) participation has a small impact on accumulated net total financial assets

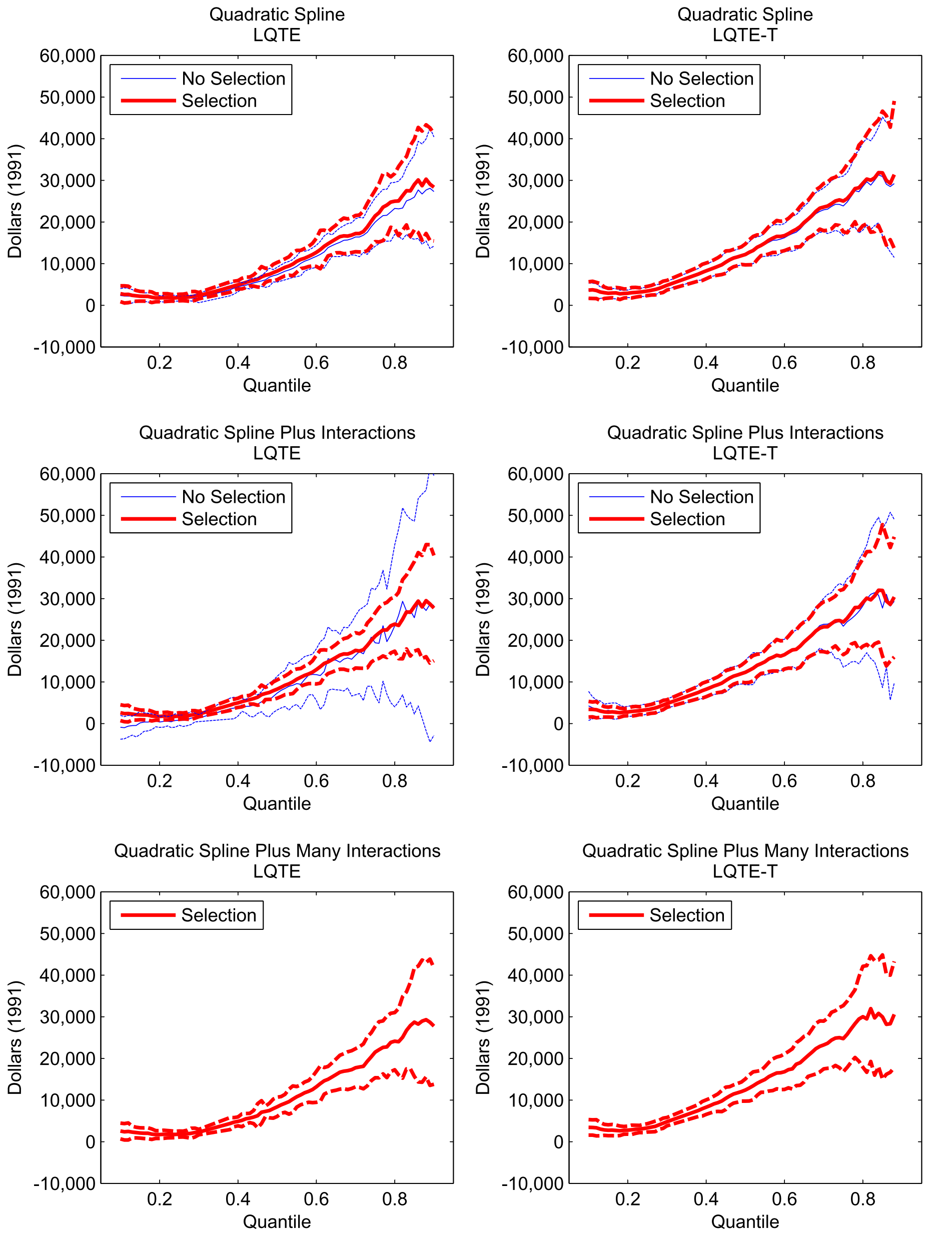

FIGURE 2.—LQTE and LQTE-T estimates of the effect of 401(k) participation on net financial assets.

at low quantiles while appearing to have a 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 the hypothesis that 401(k) participation has no effect and reject the hypothesis of a constant treatment effect more generally.

It is also worth discussing the results of the variable selection briefly as well. Due to the number of models and variable selection steps taken, especially in computing quantile effects, it is not practical to give a complete accounting of the selected variables here.

Rather, we note that for the linear model, linear IV, ATE, and LATE results, we select between two and 22 variables depending on the specification of controls and left-hand-side variable. The median number of variables selected for the QTE and LQTE results, where the median is taken across index values u, across the different specifications of controls and left-hand-side variables varies between one and 11. There is considerable variability in the number of variables selected across u, though, ranging from a minimum of no variables selected to a maximum of 237 selected variables.[28] The selected variables themselves mostly correspond to capturing the effect of income. For example, the union of the variables selected in forming each of the reduced-form quantities used for estimating the LATE in the Quadratic Spline Plus Many Interactions specification consists of 36 variables, only four of which do not include income.[29] This pattern of largely selecting terms that are direct income effects or interactions of income with other variables holds up across the specifications considered.

It is interesting that our results are similar to those in Chernozhukov and Hansen (2004) despite allowing for a much richer set of controls. 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.[30] Finally, it is worth noting that this similarity is not mechanical or otherwise built in to the procedure. For example, applications in Belloni et al. (2012) and Belloni, Chernozhukov, and Hansen (2014) use high-dimensional variable selection methods and produce sets of variables that differ substantially from intuitive baselines.

APPENDIX A: NOTATION

A.1. *Overall Notation*

We consider a random element $W = W_P$ taking values in the measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with probability law $P \in \mathcal{P}$. Note that it is most convenient to think about P as a parameter in a parameter set $\mathcal{P}$. We shall also work with a bootstrap multiplier variable ξ taking values in $(\mathbb{R}, \mathcal{A}_{\mathbb{R}})$ that is independent of W_P, having probability law P_ξ, which is fixed throughout. We consider $(W_i)_{i=1}^{\infty} = (W_{i,P})_{i=1}^{\infty}$ and $(\xi_i)_{i=1}^{\infty}$ to be i.i.d. copies of W and ξ, which are also independent of each other. The data will be defined as some measurable function of W_i for $i = 1, \dots, n$, where n denotes the sample size.

We require the sequences $(W_i)_{i=1}^{\infty}$ and $(\xi_i)_{i=1}^{\infty}$ to live on a probability space $(\Omega, \mathcal{A}_\Omega, \mathrm{P}_P)$ for all $P \in \mathcal{P}$; note that other variables arising in the proofs do not need to live on the same space. It is important to keep track of the dependence on P in the analysis since

[28]Having more than 100 variables selected occurs in the very-high-dimensional setting when the outcome in the penalized regression is $\mathbf{1}_0(D)Y_u$ for the six lowest values of u among the subset of households eligible for 401(k)'s and for the six highest values of u among the subset of households that are not eligible for 401(k)'s.

[29]Let i_1 be the indicator for income in the first income category, and define i_2–i_7 similarly. Let *db* be the defined benefit dummy, *ira* be the IRA dummy, *hown* be the home ownership dummy, *mar* be the married dummy, *te* be the two-earner household dummy, *ed* be years of schooling, and *fsize* be family size. The exact identities of the variables selected for modeling any reduced-form quantity used in estimating the LATE in the very-high-dimensional case are i_1, i_2, i_3, $income * i_3$, $income^2 * i_6$, db, $ira * hown$, $age * ira$, $ed * ira$, $i_1 * fsize$, $i_1 * fsize^2 * db$, $i_2 * fsize$, $i_2 * fsize^2 * db$, $i_3 * age^3$, $i_3 * fsize$, $i_3 * mar$, $i_3 * fsize * te$, $i_3 * fsize * mar$, $i_4 * fsize$, $i_4 * te$, $i_4 * fsize * te$, $i_4 * mar * te$, $i_4 * ed * fsize$, $i_5 * ed^2 * te$, $i_5 * fsize * te$, $income * ira$, $income * hown$, $income * mar * hown$, $income * te * hown$, $income * fsize * ira$, $income * fsize^2 * ira$, $income * i_1 * fsize$, $income^2 * ira * hown$, $income^2 * ed^2 * te$, $income^2 * i_3 * fsize$, and $income^2 * i_6 * hown$.

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

we want the results to hold uniformly in P in some set $\mathcal{P}_n$ which may be dependent on n. Typically, this set will increase with n; that is, $\mathcal{P}_n \subseteq \mathcal{P}_{n+1}$.

Throughout the paper we signify the dependence on P by mostly using P as a subscript in P_P, but in the proofs we sometimes use it as a subscript for variables, as in W_P. The operator E denotes a generic expectation operator with respect to a generic probability measure P, while E_P denotes the expectation with respect to P_P. Note also that we use capital letters such as W to denote random elements and use the corresponding lowercase letters such as w to denote fixed values that these random elements can take.

We denote by $\mathbb{P}_n$ the (random) empirical probability measure that assigns probability n^{-1} to each $W_i \in (W_i)_{i=1}^n$. $\mathbb{E}_n$ denotes the expectation with respect to the empirical measure, and $\mathbb{G}_{n,P}$ denotes the empirical process $\sqrt{n}(\mathbb{E}_n - P)$, that is,

$$\mathbb{G}_{n,P}(f) = \mathbb{G}_{n,P}\big(f(W)\big) = n^{-1/2}\sum_{i=1}^{n}\big\{f(W_i) - P\big[f(W)\big]\big\},$$

$$P\big[f(W)\big] := \int f(w)\,dP(w),$$

indexed by a measurable class of functions $\mathcal{F} : \mathcal{W} \longmapsto \mathbb{R}$; see van der Vaart and Wellner (1996, Chapter 2.3). We shall often omit the index P from $\mathbb{G}_{n,P}$ and simply write $\mathbb{G}_n$. In what follows, we use $\|\cdot\|_{\mathsf{P},q}$ to denote the $L^q(\mathsf{P})$ norm; for example, we use $\|f(W)\|_{P,q} = (\int |f(w)|^q\,dP(w))^{1/q}$ and $\|f(W)\|_{\mathbb{P}_n,q} = (n^{-1}\sum_{i=1}^n |f(W_i)|^q)^{1/q}$. For a vector $v = (v_1, \dots, v_p)' \in \mathbb{R}^p$, $\|v\|_1 = |v_1| + \cdots + |v_p|$ denotes the ℓ_1-norm of v, $\|v\| = \sqrt{v'v}$ denotes the Euclidean norm of v, and $\|v\|_0$ denotes the ℓ_0-“norm” of v which equals the number of nonzero components of v. For a positive integer k, $[k]$ denotes the set $\{1, \dots, k\}$. For x_n, y_n denoting sequences in $\mathbb{R}$, the statement $x_n \lesssim y_n$ means that $x_n \leq Ay_n$ for some constant A that does not depend on n.

We say that a collection of random variables $\mathcal{F} = \{f(W, t), t \in T\}$, where $f : \mathcal{W} \times T \to \mathbb{R}$, indexed by a set T and viewed as functions of $W \in \mathcal{W}$, is *suitably measurable* with respect to W if it is image admissible Suslin class, as defined in Dudley (1999, p. 186). In particular, $\mathcal{F}$ is suitably measurable if $f : \mathcal{W} \times T \to \mathbb{R}$ is measurable and T is a Polish space equipped with its Borel sigma algebra; see Dudley (1999, p. 186). This condition is a mild assumption satisfied in practical cases.

A.2. *Notation for Stochastic Convergence Uniformly in P*

All parameters, such as the law of the data, are indexed by P. This dependency is sometimes kept implicit. We shall allow for 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 n.

The convergence analysis, namely the 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 a sequence of stochastic processes $X_{n,P}$ and a random element Y_P, taking values in the normed space $\mathbb{D}$, defined on the probability space $(\Omega, \mathcal{A}_\Omega, \mathrm{P}_P)$. Through most of the Appendix, $\mathbb{D} = \ell^\infty(\mathcal{U})$, the space of uniformly bounded functions mapping an arbitrary index set $\mathcal{U}$ to the real line, or $\mathbb{D} = UC(\mathcal{U})$, the space of uniformly continuous functions mapping an arbitrary index set $\mathcal{U}$ to the real line. Consider also a sequence of deterministic positive constants a_n. We shall say that:

(i) $X_{n,P} = O_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\lim_{K\to\infty}\lim_{n\to\infty}\sup_{P\in\mathcal{P}_n}\mathrm{P}_P^*(|X_{n,P}| > Ka_n) = 0$,

(ii) $X_{n,P} = o_P(a_n)$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{K>0} \lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathrm{P}_P^*(|X_{n,P}| > Ka_n) = 0$,
(iii) $X_{n,P} \rightsquigarrow Y_P$ uniformly in $P \in \mathcal{P}_n$, if $\sup_{P\in\mathcal{P}_n} \sup_{h\in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_P^* h(X_{n,P}) - \mathrm{E}_P h(Y_P)| \to 0$.
Here the symbol $\rightsquigarrow$ denotes weak convergence, that is, convergence in distribution or law, $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functions mapping $\mathbb{D}$ to $[0, 1]$ with Lipschitz norm at most 1, and the outer probability and expectation, P_P^* and E_P^*, are invoked whenever (non)-measurability arises.

LEMMA A.1: *The above notions* (i), (ii), *and* (iii) *are equivalent to the following notions* (a), (b), *and* (c), *each holding for every sequence* $P_n \in \mathcal{P}_n$:
(a) $X_{n,P_n} = O_{P_n}(a_n)$, *that is*, $\lim_{K\to\infty} \lim_{n\to\infty} \mathrm{P}_{P_n}^*(|X_{n,P_n}| > Ka_n) = 0$;
(b) $X_{n,P_n} = o_{P_n}(a_n)$, *that is*, $\sup_{K>0} \lim_{n\to\infty} \mathrm{P}_{P_n}^*(|X_{n,P_n}| > Ka_n) = 0$;
(c) $X_{n,P_n} \rightsquigarrow Y_{P_n}$, *that is*, $\sup_{h\in\mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_{P_n}^* h(X_{n,P_n}) - \mathrm{E}_{P_n} h(Y_{P_n})| \to 0$.

The claims follow straightforwardly from the definitions, so the proof is omitted. We shall use this equivalence extensively in the proofs of the main results without explicit reference.

APPENDIX B: KEY TOOLS I: UNIFORM IN P DONSKER THEOREM, MULTIPLIER BOOTSTRAP, AND FUNCTIONAL DELTA METHOD

B.1. *Uniform in P Donsker Property*

Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of the random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ according to the probability law P on that space. Let $\mathcal{F}_P = \{f_{t,P} : t \in T\}$ be a set of suitably measurable functions $w \longmapsto f_{t,P}(w)$ mapping $\mathcal{W}$ to $\mathbb{R}$, equipped with a measurable envelope $F_P : \mathcal{W} \longmapsto \mathbb{R}$. The class is indexed by $P \in \mathcal{P}$ and $t \in T$, where T is a fixed, totally bounded semimetric space equipped with a semimetric d_T. Let $N(\epsilon, \mathcal{F}_P, \|\cdot\|_{Q,2})$ denote the ϵ-covering number of the class of functions $\mathcal{F}_P$ with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$ for Q a finitely-discrete measure on $(\mathcal{W}, \mathcal{A}_\mathcal{W})$. We shall use the following result.

THEOREM B.1—Uniform in P Donsker Property: *Work with the setup above. Suppose that, for $q > 2$,*

$$\sup_{P\in\mathcal{P}} \|F_P\|_{P,q} \le C \quad \textit{and} \quad \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \sup_{d_T(t,\bar{t})\le\delta} \|f_{t,P} - f_{\bar{t},P}\|_{P,2} = 0. \tag{B.1}$$

Furthermore, suppose that

$$\lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \int_0^\delta \sup_Q \sqrt{\log N\big(\epsilon\|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}\big)}\, d\epsilon = 0. \tag{B.2}$$

Let $\mathbb{G}_P$ denote the P-Brownian bridge, and consider

$$Z_{n,P} := \big(Z_{n,P}(t)\big)_{t\in T} := \big(\mathbb{G}_n(f_{t,P})\big)_{t\in T},$$
$$Z_P := \big(Z_P(t)\big)_{t\in T} := \big(\mathbb{G}_P(f_{t,P})\big)_{t\in T}.$$

(a) *Then, $Z_{n,P} \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely,*

$$\sup_{P\in\mathcal{P}} \sup_{h\in\mathrm{BL}_1(\ell^\infty(T))} \big|\mathrm{E}_P^* h(Z_{n,P}) - \mathrm{E}_P h(Z_P)\big| \to 0.$$

(b) *The process* $Z_{n,P}$ *is stochastically equicontinuous uniformly in* $P \in \mathcal{P}$, *that is*, *for every* $\varepsilon > 0$,

$$\lim_{\delta \searrow 0} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathrm{P}_P^*\Big(\sup_{d_T(t,\bar{t})\le\delta} \big|Z_{n,P}(t) - Z_{n,P}(\bar{t})\big| > \varepsilon\Big) = 0.$$

(c) *The limit process* Z_P *has the following continuity properties*:

$$\sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{t\in T}\big|Z_P(t)\big| < \infty, \quad \lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \mathrm{E}_P \sup_{d_T(t,\bar{t})\le\delta} \big|Z_P(t) - Z_P(\bar{t})\big| = 0.$$

(d) *The paths* $t \longmapsto Z_P(t)$ *are a.s. uniformly continuous on* (T, d_T) *under each* $P \in \mathcal{P}$.

COMMENT B.1—Important Feature of the Theorem: This is an extension of the uniform Donsker theorem stated in Theorem 2.8.2 in van der Vaart and Wellner (1996), which allows for the function classes $\mathcal{F}$ to be *dependent on* P. This generalization is crucial and is required in all of our problems.

B.2. *Uniform in P Validity of Multiplier Bootstrap*

Consider the setting of the preceding subsection. Let $(\xi_i)_{i=1}^n$ be i.i.d. multipliers whose distribution does not depend on P, such that $\mathrm{E}\xi = 0$, $\mathrm{E}\xi^2 = 1$, and $\mathrm{E}|\xi|^q \le C$ for $q > 2$. Consider the multiplier empirical process:

$$Z_{n,P}^* := \big(Z_{n,P}^*(t)\big)_{t\in T} := \big(\mathbb{G}_n(\xi f_{t,P})\big)_{t\in T} := \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f_{t,P}(W_i)\right)_{t\in T}.$$

Here $\mathbb{G}_n$ is taken to be an extended empirical process defined by the empirical measure that assigns mass $1/n$ to each point (W_i, ξ_i) for $i = 1, \dots, n$. Let $Z_P = (Z_P(t))_{t\in T} = (\mathbb{G}_P(f_{t,P}))_{t\in T}$ as defined in Theorem B.1.

THEOREM B.2—Uniform in P Validity of Multiplier Bootstrap: *Assume the conditions of Theorem* B.1 *hold*. *Then* (a) *the following unconditional convergence takes place*, $Z_{n,P}^* \rightsquigarrow Z_P$ *in* $\ell^\infty(T)$ *uniformly in* $P \in \mathcal{P}$, *namely*,

$$\sup_{P\in\mathcal{P}} \sup_{h\in \mathrm{BL}_1(\ell^\infty(T))} \big|\mathrm{E}_P^* h\big(Z_{n,P}^*\big) - \mathrm{E}_P h(Z_P)\big| \to 0,$$

and (b) *the following conditional convergence takes place*, $Z_{n,P}^* \rightsquigarrow_B Z_P$ *in* $\ell^\infty(T)$ *uniformly in* $P \in \mathcal{P}$, *namely*, *uniformly in* $P \in \mathcal{P}$,

$$\sup_{h\in \mathrm{BL}_1(\ell^\infty(T))} \big|\mathrm{E}_{B_n} h\big(Z_{n,P}^*\big) - \mathrm{E}_P h(Z_P)\big| = o_P^*(1),$$

where E_{B_n} *denotes the expectation over the multiplier weights* $(\xi_i)_{i=1}^n$ *holding the data* $(W_i)_{i=1}^n$ *fixed*.

B.3. *Uniform in P Functional Delta Method and Bootstrap*

We shall use the functional delta method, as formulated in van der Vaart and Wellner (1996, Chapter 3.9). Let $\mathbb{D}_0$, $\mathbb{D}$, and $\mathbb{E}$ be normed spaces, with $\mathbb{D}_0 \subset \mathbb{D}$. A map

$\phi : \mathbb{D}_\phi \subset \mathbb{D} \longmapsto \mathbb{E}$ is called *Hadamard-differentiable* at $\rho \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ if there is a continuous linear map $\phi'_\rho : \mathbb{D}_0 \longmapsto \mathbb{E}$ such that

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

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

We now define the following notion of the uniform Hadamard differentiability:

DEFINITION B.1—Uniform Hadamard Tangential Differentiability: Consider a map $\phi : \mathbb{D}_\phi \longmapsto \mathbb{E}$, where the domain of the map $\mathbb{D}_\phi$ is a subset of a normed space $\mathbb{D}$ and the range is a subset of the normed space $\mathbb{E}$. Let $\mathbb{D}_0$ be a normed space, with $\mathbb{D}_0 \subset \mathbb{D}$, and $\mathbb{D}_\rho$ be a compact metric space, a subset of $\mathbb{D}_\phi$. The map $\phi : \mathbb{D}_\phi \longmapsto \mathbb{E}$ is called *Hadamard-differentiable uniformly* in $\rho \in \mathbb{D}_\rho$ tangentially to $\mathbb{D}_0$ with derivative map $h \longmapsto \phi'_\rho(h)$, if

$$\left| \frac{\phi(\rho_n + t_n h_n) - \phi(\rho_n)}{t_n} - \phi'_\rho(h) \right| \to 0,$$

$$\left| \phi'_{\rho_n}(h_n) - \phi'_\rho(h) \right| \to 0, \quad n \to \infty,$$

for all convergent sequences $\rho_n \to \rho$ in $\mathbb{D}_\rho$, $t_n \to 0$ in $\mathbb{R}$, and $h_n \to h \in \mathbb{D}_0$ in $\mathbb{D}$ such that $\rho_n + t_n h_n \in \mathbb{D}_\phi$ for every n. As a part of the definition, we require that the derivative map $h \longmapsto \phi'_\rho(h)$ from $\mathbb{D}_0$ to $\mathbb{E}$ is linear for each $\rho \in \mathbb{D}_\rho$.

COMMENT B.2: Note that the definition requires that the derivative map $(\rho, h) \longmapsto \phi'_\rho(h)$, mapping $\mathbb{D}_\rho \times \mathbb{D}_0$ to $\mathbb{E}$, is continuous at each $(\rho, h) \in \mathbb{D}_\rho \times \mathbb{D}_0$.

COMMENT B.3—Important Details of the Definition: Definition B.1 is different from the definition of uniform differentiability given in van der Vaart and Wellner (1996, p. 379, Equation (3.9.12)), since our definition allows $\mathbb{D}_\rho$ to be much smaller than $\mathbb{D}_\phi$ and allows $\mathbb{D}_\rho$ to be endowed with a much stronger metric than the metric induced by the norm of $\mathbb{D}$. These differences are essential for infinite-dimensional applications. For example, the quantile/inverse map is uniformly Hadamard-differentiable in the sense of Definition B.1 for a suitable choice of $\mathbb{D}_\rho$: Let $T = [\epsilon, 1 - \epsilon]$, $\mathbb{D} = \ell^\infty(T)$, $\mathbb{D}_\phi =$ set of càdlàg functions on T, $\mathbb{D}_0 = \mathrm{UC}(T)$, and $\mathbb{D}_\rho$ be a compact subset of $C^1(T)$ such that each $\rho \in \mathbb{D}_\rho$ obeys $\partial\rho(t)/\partial t \geq c > 0$ on $t \in T$, where c is a positive constant. However, the quantile/inverse map is not Hadamard-differentiable uniformly on $\mathbb{D}_\rho$ if we set $\mathbb{D}_\rho = \mathbb{D}_\phi$ and hence is not uniformly differentiable in the sense of the definition given in van der Vaart and Wellner (1996) which requires $\mathbb{D}_\rho = \mathbb{D}_\phi$. It is important and practical to keep the distinction between $\mathbb{D}_\rho$ and $\mathbb{D}_\phi$ since the estimated values $\hat\rho$ may well be outside $\mathbb{D}_\rho$ unless explicitly imposed in estimation even though the population values of ρ are in $\mathbb{D}_\rho$ by assumption. For example, the empirical c.d.f. is in $\mathbb{D}_\phi$, but is outside $\mathbb{D}_\rho$.

THEOREM B.3—Functional Delta Method Uniformly in $P \in \mathcal{P}$: *Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \longmapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\rho \in \mathbb{D}_\rho \subset \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ with derivative map ϕ'_ρ. Let $\hat\rho_{n,P}$ be a sequence of stochastic processes taking values in $\mathbb{D}_\phi$, where each $\hat\rho_{n,P}$ is an estimator of the parameter $\rho_P \in \mathbb{D}_\rho$. Suppose there exists a sequence of constants $r_n \to \infty$ such that $Z_{n,P} = r_n(\hat\rho_{n,P} - \rho_P) \rightsquigarrow Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. The limit process Z_P is separable and takes its values in $\mathbb{D}_0$ for all $P \in \mathcal{P} = \bigcup_{n \geq n_0} \mathcal{P}_n$, where n_0 is fixed. Moreover, the set of stochastic processes $\{Z_P : P \in \mathcal{P}\}$ is relatively compact in the topology of weak convergence* *in $\mathbb{D}_0$, that is, every sequence in this set can be split into weakly convergent subsequences. Then, $r_n(\phi(\hat\rho_{n,P}) - \phi(\rho_P)) \rightsquigarrow \phi'_{\rho_P}(Z_P)$ in $\mathbb{E}$ uniformly in $P \in \mathcal{P}_n$. If $(\rho, h) \longmapsto \phi'_\rho(h)$ is defined and continuous on the whole of $\mathbb{D}_\rho \times \mathbb{D}$, then the sequence $r_n(\phi(\hat\rho_{n,P}) - \phi(\rho_P)) - \phi'_{\rho_P}(r_n(\hat\rho_{n,P} - \rho_P))$ converges to zero in outer probability uniformly in $P \in \mathcal{P}_n$. Moreover, the set of stochastic processes $\{\phi'_{\rho_P}(Z_P) : P \in \mathcal{P}\}$ is relatively compact in the topology of weak convergence in $\mathbb{E}$.*

The following result on the functional delta method applies to any bootstrap or other simulation method obeying certain conditions. Such methods include the multiplier bootstrap as a special case. Let $D_{n,P} = (W_{i,P})_{i=1}^n$ denote the data vector and $B_n = (\xi_i)_{i=1}^n$ be a vector of random variables used to generate bootstrap or simulation draws (the specifics may vary depending on the particular method employed). Consider sequences of stochastic processes $\hat\rho_{n,P} = \hat\rho_{n,P}(D_{n,P})$, where $Z_{n,P} = r_n(\hat\rho_{n,P} - \rho_P) \rightsquigarrow Z_P$ in the normed space $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. Also consider the bootstrap stochastic process $Z^*_{n,P} = Z_{n,P}(D_{n,P}, B_n)$ in $\mathbb{D}$, where $Z_{n,P}$ is a measurable function of B_n for each value of D_n. Suppose that $Z^*_{n,P}$ converges conditionally given D_n in distribution to Z_P uniformly in $P \in \mathcal{P}_n$, namely that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D})} \left| \mathrm{E}_{B_n}\left[h\left(Z^*_{n,P}\right)\right] - \mathrm{E}_P h(Z_P)\right| = o^*_P(1),$$

uniformly in $P \in \mathcal{P}_n$, where E_{B_n} denotes the expectation computed with respect to the law of B_n holding the data $D_{n,P}$ fixed. This is denoted as "$Z^*_{n,P} \rightsquigarrow_B Z_P$ uniformly in $P \in \mathcal{P}_n$." Finally, let $\hat\rho^*_{n,P} = \hat\rho_{n,P} + Z^*_{n,P}/r_n$ denote the bootstrap or simulation draw of $\hat\rho_{n,P}$.

THEOREM B.4—Uniform in P Functional Delta Method for Bootstrap and Other Simulation Methods: *Assume the conditions of Theorem B.3 hold. Let $\hat\rho_{n,P}$ and $\hat\rho^*_{n,P}$ be maps as indicated previously taking values in $\mathbb{D}_\phi$ such that $r_n(\hat\rho_{n,P} - \rho_P) \rightsquigarrow Z_P$ and $r_n(\hat\rho^*_{n,P} - \hat\rho_{n,P}) \rightsquigarrow_B Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. Then, $X^*_{n,P} = r_n(\phi(\hat\rho^*_{n,P}) - \phi(\hat\rho_{n,P})) \rightsquigarrow_B X_P = \phi'_{\rho_P}(Z_P)$ uniformly in $P \in \mathcal{P}_n$.*

B.4. *Proof of Theorem B.1*

Parts (a) and (b) are a direct consequence of Lemma B.2. In particular, Lemma B.2(a) implies stochastic equicontinuity under arbitrary subsequences $P_n \in \mathcal{P}$, which implies part (b). Part (a) follows from Lemma B.2(b) by splitting an arbitrary sequence $n \in \mathbb{N}$ into subsequences $n \in \mathbb{N}'$ along each of which the covariance function $(t, s) \longmapsto c_{P_n}(t, s) := P_n f_{s,P_n} f_{t,P_n} - P_n f_{s,P_n} P_n f_{t,P_n}$ converges uniformly and therefore also pointwise to a uniformly continuous function on (T, d_T). This convergence is possible because $\{(t, s) \longmapsto c_P(t, s) : P \in \mathcal{P}\}$ is a relatively compact set in $\ell^\infty(T \times T)$ in view of the Arzela–Ascoli theorem, the assumptions in equation (B.1), and total boundedness of (T, d_T). By Lemma B.2(b), pointwise convergence of the covariance function implies weak convergence to a tight Gaussian process which may depend on the identity $\mathbb{N}'$ of the subsequence. Since this argument applies to each such subsequence that split the overall sequence, part (b) follows.

Part (c) is immediate from the imposed uniform covering entropy condition and Dudley's metric entropy inequality for expectations of suprema of Gaussian processes (Corollary 2.2.8 in van der Vaart and Wellner (1996)). Claim (d) follows from claim (c) and a standard argument, based on the application of the Borel–Cantelli lemma. Indeed, let $m \in \mathbb{N}$ be a sequence and $\delta_m := 2^{-m} \wedge \sup\{\delta > 0 : \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_T(t,\bar t) \le \delta} |Z_P(t) - Z_P(\bar t)| < 2^{-2m}\}$; then by the Markov inequality, $\mathrm{P}_P(\sup_{d_T(t,\bar t) \le \delta_m} |Z_P(t) - Z_P(\bar t)| > 2^{-m}) \le 2^{-2m+m} = 2^{-m}$. This sums to a finite number over $m \in \mathbb{N}$. Hence, by the Borel–Cantelli lemma, for almost all states $\omega \in \Omega$, $|Z_P(t)(\omega) - Z_P(\bar{t})(\omega)| \le 2^{-m}$ for all $d_T(t, \bar{t}) \le \delta_m \le 2^{-m}$ and all m sufficiently large. Hence claim (d) follows. *Q.E.D.*

B.5. *Proof of Theorem B.2*

Claim (a) is verified by invoking Theorem B.1. We begin by showing that $Z_P^* = (\mathbb{G}_P \xi f_{t,P})_{t\in T}$ is equal in distribution to $Z_P = (\mathbb{G}_P f_{t,P})_{t\in T}$; in particular, Z_P^* and Z_P share identical mean and covariance function, and thus they share the continuity properties established in Theorem B.1. This claim is immediate from the fact that multiplication by ξ of each $f \in \mathcal{F}_P = \{f_{t,P} : t \in T\}$ yields a set $\xi\mathcal{F}_P$ of measurable functions $\xi f : (w, \xi) \longmapsto \xi f(w)$, mapping $\mathcal{W} \times \mathbb{R}$ to $\mathbb{R}$. Each such function has mean zero under $P \times P_\xi$, that is, $\int sf(w)\,dP_\xi(s)\,dP(w) = 0$, and covariance function $(\xi f, \xi\tilde{f}) \longmapsto Pf\tilde{f} - PfP\tilde{f}$. Hence the Gaussian process $(\mathbb{G}_P(\xi f))_{\xi f\in\xi\mathcal{F}_P}$ shares the zero mean and the covariance function of $(\mathbb{G}_P(f))_{f\in\mathcal{F}_P}$.

We are claiming that $Z_{n,P}^* \rightsquigarrow Z_P^*$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, where $Z_{n,P}^* := (\mathbb{G}_n \xi f_{t,P})_{t\in T}$. We note that the function class $\mathcal{F}_P$ and the corresponding envelope F_P satisfy the conditions of Theorem B.1. The same is also true for the function class $\xi\mathcal{F}_P$ defined by $(w, \xi) \longmapsto \xi f_P(w)$, which maps $\mathcal{W} \times \mathbb{R}$ to $\mathbb{R}$ and its envelope $|\xi|F_P$, since ξ is independent of W. Let Q now denote a finitely discrete measure over $\mathcal{W} \times \mathbb{R}$. By Lemma K.1 multiplication by ξ does not change qualitatively the uniform covering entropy bound:

$$\log \sup_Q N\big(\epsilon \big\| |\xi| F_P \big\|_{Q,2}, \xi\mathcal{F}_P, \|\cdot\|_{Q,2}\big)$$
$$\le \log \sup_Q N\big(2^{-1}\epsilon \|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}\big).$$

Moreover, multiplication by ξ does not affect the norms, $\|\xi f_P(W)\|_{P\times P_\xi,2} = \|f_P(W)\|_{P,2}$, since ξ is independent of W by construction and $\mathrm{E}\xi^2 = 1$. The claim then follows.

Claim (b). For each $\delta > 0$ and $t \in T$, let $\pi_\delta(t)$ denote a closest element in a given, finite δ-net over T. We begin by noting that

$$\begin{aligned}
\Delta_P &:= \sup_{h\in \mathrm{BL}_1} \big|\mathrm{E}_{B_n} h\big(Z_{n,P}^*\big) - \mathrm{E}_P h(Z_P)\big| \\
&\le \mathrm{I}_P + \mathrm{II}_P + \mathrm{III}_P \\
&:= \sup_{h\in \mathrm{BL}_1} \big|\mathrm{E}_P h(Z_P \circ \pi_\delta) - \mathrm{E}_P h(Z_P)\big| \\
&\quad + \sup_{h\in \mathrm{BL}_1} \big|\mathrm{E}_{B_n} h\big(Z_{n,P}^* \circ \pi_\delta\big) - \mathrm{E}_P h(Z_P \circ \pi_\delta)\big| \\
&\quad + \sup_{h\in \mathrm{BL}_1} \big|\mathrm{E}_{B_n} h\big(Z_{n,P}^* \circ \pi_\delta\big) - \mathrm{E}_{B_n} h\big(Z_{n,P}^*\big)\big|,
\end{aligned}$$

where here and below BL_1 abbreviates $\mathrm{BL}_1(\ell^\infty(T))$.

First, we note that $\mathrm{I}_P \le \mathrm{E}_P(\sup_{d_T(t,\bar{t})\le\delta} |Z_P(t) - Z_P(\bar{t})| \wedge 2) =: \mu_P(\delta)$ and $\lim_{\delta\searrow 0} \sup_{P\in\mathcal{P}} \mu_P(\delta) = 0$. The first assertion follows from

$$\begin{aligned}
\mathrm{I}_P &\le \sup_{h\in \mathrm{BL}_1} \mathrm{E}_P \big|h\big(Z_{n,P}^* \circ \pi_\delta\big) - h\big(Z_{n,P}^*\big)\big| \\
&\le \mathrm{E}_P\Big(\sup_{t\in T} \big|Z_P \circ \pi_\delta(t) - Z_P(t)\big| \wedge 2\Big) \le \mu_P(\delta),
\end{aligned}$$

and the second assertion holds by Theorem B.1(c).

Second, $\mathrm{E}_P^*\mathrm{III}_P \leq \mathrm{E}_P^*(\sup_{d_T(t,\bar{t})\leq\delta}|Z_{n,P}^*(t) - Z_{n,P}^*(\bar{t})| \wedge 2) =: \mu_P^*(\delta)$ and $\lim_{n\to\infty}\sup_{P\in\mathcal{P}}|\mu_P^*(\delta)-\mu_P(\delta)|=0$. The first assertion follows because $\mathrm{E}_P^*\mathrm{III}_P$ is bounded by

$$\mathrm{E}_P^* \sup_{h\in\mathrm{BL}_1} \mathrm{E}_{B_n}\big|h\big(Z_{n,P}^*\circ\pi_\delta\big)-h\big(Z_{n,P}^*\big)\big|$$
$$\leq \mathrm{E}_P^*\mathrm{E}_{B_n}\Big(\sup_{t\in T}\big|Z_{n,P}^*\circ\pi_\delta(t)-Z_{n,P}^*(t)\big|\wedge 2\Big)\leq\mu_P^*(\delta).$$

The second assertion holds by part (a) of the present theorem.

Define $\epsilon(\delta):=\delta\vee\sup_{P\in\mathcal{P}}\mu_P(\delta)$. Then, by Markov's inequality, followed by $n\to\infty$,

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\mathrm{P}_P^*\big(\mathrm{III}_P>\sqrt{\epsilon(\delta)}\big)\leq\limsup_{n\to\infty}\frac{\sup_{P\in\mathcal{P}}\mu_P^*(\delta)}{\sqrt{\epsilon(\delta)}}$$
$$\leq\frac{\sup_{P\in\mathcal{P}}\mu_P(\delta)}{\sqrt{\epsilon(\delta)}}\leq\sqrt{\epsilon(\delta)}.$$

Finally, by Lemma B.1, for each $\varepsilon>0$, $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\mathrm{P}_P^*(\mathrm{II}_P>\varepsilon)=0$.

We can now conclude. Note that $\epsilon(\delta)\searrow 0$ if $\delta\searrow 0$, which holds by the definition of $\epsilon(\delta)$ and the property $\sup_{P\in\mathcal{P}}\mu_P(\delta)\searrow 0$ if $\delta\searrow 0$ noted above. Hence for each $\varepsilon>0$ and all $0<\delta<\delta_\varepsilon$ such that $3\sqrt{\epsilon(\delta)}<\varepsilon$,

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\mathrm{P}_P^*(\varDelta_P>\varepsilon)\leq\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\mathrm{P}_P^*\big(\mathrm{I}_P+\mathrm{II}_P+\mathrm{III}_P>3\sqrt{\epsilon(\delta)}\big)$$
$$\leq\sqrt{\epsilon(\delta)}.$$

Sending $\delta\searrow 0$ gives the result. *Q.E.D.*

B.6. *Auxiliary Result: Conditional Multiplier CLT in $\mathbb{R}^d$ Uniformly in $P\in\mathcal{P}$*

We rely on the following lemma, which is apparently new. An analogous result can be derived for almost sure convergence from well-known non-uniform multiplier central limit theorems, but this strategy requires us to put all the variables indexed by P on the single underlying probability space, which is much less convenient in applications.

LEMMA B.1—Conditional Multiplier Central Limit Theorem in $\mathbb{R}^d$ Uniformly in $P\in\mathcal{P}$: *Let $(Z_{i,P})_{i=1}^\infty$ be i.i.d. random vectors on $\mathbb{R}^d$, indexed by a parameter $P\in\mathcal{P}$. The parameter P represents probability laws on $\mathbb{R}^d$. For each $P\in\mathcal{P}$, these vectors are assumed to be independent of the i.i.d. sequence $(\xi_i)_{i=1}^\infty$ with $\mathrm{E}[\xi]=0$ and $\mathrm{E}[\xi^2]=1$. There exist constants $2<q<\infty$ and $0<M<\infty$, such that $\mathrm{E}_PZ_{1,P}=0$ and $(\mathrm{E}_P\|Z_{1,P}\|^q)^{1/q}\leq M$ uniformly for all $P\in\mathcal{P}$. Then, for every $\varepsilon>0$,*

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}}\mathrm{P}_P^*\Bigg(\sup_{h\in\mathrm{BL}_1(\mathbb{R}^d)}\Bigg|\mathrm{E}_{B_n}h\Bigg(n^{-1/2}\sum_{i=1}^n\xi_iZ_{i,P}\Bigg)$$
$$-\mathrm{E}_Ph\big(N\big(0,\mathrm{E}_PZ_{1,P}Z_{1,P}'\big)\big)\Bigg|>\varepsilon\Bigg)=0,$$

where E_{B_n} denotes the expectation over $(\xi_i)_{i=1}^n$ holding $(Z_{i,P})_{i=1}^n$ fixed.

PROOF: Let X and Y be random variables in $\mathbb{R}^d$, then define $d_{\mathrm{BL}}(X, Y) := \sup_{h\in \mathrm{BL}_1(\mathbb{R}^d)} |\mathrm{E}h(X) - \mathrm{E}h(Y)|$. It suffices to show that for any sequence $P_n \in \mathcal{P}$ and $N^* \sim n^{-1/2}\sum_{i=1}^n \xi_i Z_{i,P_n} | (Z_{i,P_n})_{i=1}^n$, $d_{\mathrm{BL}}(N^*, N(0, \mathrm{E}_{P_n} Z_{1,P_n} Z'_{1,P_n})) \to 0$ in probability (under P_{P_n}).

Following Bickel and Freedman (1981), we shall rely on the Mallow's metric, written m_r, which is a metric on the space of distribution functions on $\mathbb{R}^d$. For our purposes it suffices to recall that, given a sequence of distribution functions $\{F_k\}$ and a distribution function F, $m_r(F_k, F) \to 0$ if and only if $\int g\, dF_k \to \int g\, dF$ for each continuous and bounded $g : \mathbb{R}^d \to \mathbb{R}$, and $\int \|z\|^r\, dF_k(z) \to \int \|z\|^r\, dF(z)$. See Bickel and Freedman (1981) for the definition of m_r.

Under the assumptions of the lemma, we can split the sequence $n \in \mathbb{N}$ into subsequences $n \in \mathbb{N}'$, along each of which the distribution function of Z_{1,P_n} converges to some distribution function F' with respect to the Mallow's metric m_r, for some $2 < r < q$. This also implies that $N(0, \mathrm{E}_{P_n} Z_{1,P_n} Z'_{1,P_n})$ converges weakly to a normal limit $N(0, Q')$ with $Q' = \int zz'\, dF'(z)$ such that $\|Q'\| \leq M$. Both Q' and F' can depend on the subsequence $\mathbb{N}'$.

Let F_k be the empirical distribution function of a sequence $(z_i)_{i=1}^k$ of constant vectors in $\mathbb{R}^d$, where $k \in \mathbb{N}$. The law of $N^*_{F_k} = k^{-1/2}\sum_{i=1}^k \xi_i z_i$ is completely determined by F_k and the law of ξ (the latter is fixed, so it does not enter as the subscript in the definition of $N^*_{F_k}$). If $m_r(F_k, F') \to 0$ as $k \to \infty$, then $d_{\mathrm{BL}}(N^*_{F_k}, N(0, Q')) \to 0$ by Lindeberg's central limit theorem.

Let $\mathbb{F}_n$ denote the empirical distribution function of $(Z_{i,P_n})_{i=1}^n$. Note that $N^* = N^*_{\mathbb{F}_n} \sim n^{-1/2}\sum_{i=1}^n \xi_i Z_{i,P_n} | (Z_{i,P_n})_{i=1}^n$. By the law of large numbers for arrays, $\int g\, d\mathbb{F}_n \to \int g\, dF'$ and $\int \|z\|^r\, d\mathbb{F}_n(z) \to \int \|z\|^r\, dF'(z)$ in probability along the subsequence $n \in \mathbb{N}'$. Hence $m_r(\mathbb{F}_n, F') \to 0$ in probability along the same subsequence. We can conclude that $d_{\mathrm{BL}}(N^*_{\mathbb{F}_n}, N(0, Q')) \to 0$ in probability along the same subsequence by the extended continuous mapping theorem (van der Vaart and Wellner (1996, Theorem 1.11.1)).

The argument applies to every subsequence $\mathbb{N}'$ of the stated form. The claim in the first paragraph of the proof thus follows. *Q.E.D.*

B.7. *Donsker Theorems for Function Classes That Depend on n*

Let $(W_i)_{i=1}^\infty$ be a sequence of i.i.d. copies of the random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$, whose law is determined by the probability measure P, and let $w \longmapsto f_{n,t}(w)$ be measurable functions $f_{n,t} : \mathcal{W} \to \mathbb{R}$ indexed by $n \in \mathbb{N}$ and a fixed, totally bounded semimetric space (T, d_T). Consider the stochastic process

$$(\mathbb{G}_n f_{n,t})_{t\in T} := \left\{ n^{-1/2} \sum_{i=1}^n \big(f_{n,t}(W_i) - P f_{n,t}\big) \right\}_{t\in T}.$$

This empirical process is indexed by a class of functions $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ with a measurable envelope function F_n. It is important to note here that the dependence on n allows us to have *the class itself* be possibly dependent on the law P_n.

LEMMA B.2—Donsker Theorem for Classes Changing With n: *Work with the setup above. Suppose that for some fixed constant $q > 2$ and every sequence $\delta_n \searrow 0$,*

$$\|F_n\|_{P_n,q} = O(1), \qquad \sup_{d_T(s,t)\leq \delta_n} \|f_{n,s} - f_{n,t}\|_{P_n,2} \to 0,$$

$$\int_0^{\delta_n} \sup_Q \sqrt{\log N\big(\epsilon \|F_n\|_{Q,2}, \mathcal{F}_n, \|\cdot\|_{Q,2}\big)}\, d\epsilon \to 0.$$

(a) *Then the empirical process* $(\mathbb{G}_n f_{n,t})_{t\in T}$ *is asymptotically tight in* $\ell^\infty(T)$, *that is, stochastically equicontinuous*. (b) *For any subsequence such that the covariance function* $P_n f_{n,s} f_{n,t} - P_n f_{n,s} P_n f_{n,t}$ *converges pointwise on* $T \times T$, $(\mathbb{G}_n f_{n,t})_{t\in T}$ *converges in* $\ell^\infty(T)$ *to a Gaussian process with covariance function given by the limit of the covariance function along that subsequence*.

PROOF: The use of Theorem 2.11.1 in van der Vaart and Wellner (1996), which does allow for the probability space to depend on n, allows us to establish claim (a), by repeating the proof (verbatim) of Theorem 2.11.22 in van der Vaart and Wellner (1996, pp. 220–221), except that the probability law is allowed to depend on n. (For the sake of completeness, the Supplemental Material provides the complete proof.) The proof of claim (b) follows by a standard argument from the stochastic equicontinuity established in claim (a) and finite-dimensional convergence along the indicated subsequences. *Q.E.D.*

B.8. *Proof of Theorems B.3 and B.4*

The proof consists of two parts, each proving the corresponding theorem.

Part 1. We can split $\mathbb{N}$ into subsequences $\{\mathbb{N}'\}$ along each of which $Z_{n,P_n} \rightsquigarrow Z' \in \mathbb{D}_0$ in $\mathbb{D}$, $\rho_{P_n} \to \rho'$ in $\mathbb{D}_\rho$ $(n \in \mathbb{N}')$, where Z' and ρ' can possibly depend on $\mathbb{N}'$. It suffices to verify that, for each $\mathbb{N}'$,

$$r_n\big(\phi(\hat\rho_{n,P_n}) - \phi(\rho_{P_n})\big) \rightsquigarrow \phi'_{\rho'}\big(Z'\big) \quad \big(n \in \mathbb{N}'\big), \tag{B.3}$$

$$r_n\big(\phi(\hat\rho_{n,P_n}) - \phi(\rho_{P_n})\big) - \phi'_{\rho_{P_n}}\big(r_n(\hat\rho_{n,P_n} - \rho_{P_n})\big) \rightsquigarrow 0 \quad \big(n \in \mathbb{N}'\big), \tag{B.4}$$

$$r_n\big(\phi(\hat\rho_{n,P_n}) - \phi(\rho_{P_n})\big) - \phi'_{\rho'}\big(r_n(\hat\rho_{n,P_n} - \rho_{P_n})\big) \rightsquigarrow 0 \quad \big(n \in \mathbb{N}'\big), \tag{B.5}$$

where the last two claims hold provided that $(\rho, h) \longmapsto \phi'_\rho(h)$ is defined and continuous on the whole of $\mathbb{D}_\rho \times \mathbb{D}$. The claim (B.5) is not needed in Part 1, but we need it for Part 2.

The map $g_n(h) = r_n(\phi(\rho_{P_n} + r_n^{-1}h) - \phi(\rho_{P_n}))$, from $\mathbb{D}_n = \{h \in \mathbb{D} : \rho_{P_n} + r_n^{-1}h \in \mathbb{D}_\phi\}$ to $\mathbb{E}$, satisfies $g_n(h_n) \to \phi'_{\rho'}(h)$ for every subsequence $h_n \to h \in \mathbb{D}_0$ (with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner (1996, Theorem 1.11.1)) yields (B.3).

Similarly, the map $m_n(h) = r_n(\phi(\rho_{P_n} + r_n^{-1}h) - \phi(\rho_{P_n})) - \phi'_{\rho_{P_n}}(h)$, from $\mathbb{D}_n = \{h \in \mathbb{D} : \rho_{P_n} + r_n^{-1}h \in \mathbb{D}_\phi\}$ to $\mathbb{E}$, satisfies $m_n(h_n) \to \phi'_{\rho'}(h) - \phi'_{\rho'}(h) = 0$ for every subsequence $h_n \to h \in \mathbb{D}_0$ (with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner (1996, Theorem 1.11.1)) yields (B.4). The proof of (B.5) is completely analogous and is omitted.

To establish relative compactness, we work with each $\mathbb{N}'$. Then $\phi'_{\rho_{P_n}}(h)$ mapping $\mathbb{D}_0$ to $\mathbb{E}$ satisfies $\phi'_{\rho_{P_n}}(h_n) \to \phi'_{\rho'}(h)$ for every subsequence $h_n \to h \in \mathbb{D}_0$ (with $n \in \mathbb{N}'$). Application of the extended continuous mapping theorem (van der Vaart and Wellner (1996, Theorem 1.11.1)) yields that $\phi'_{\rho_{P_n}}(Z_P) \rightsquigarrow \phi'_{\rho'}(Z')$.

Part 2. We can split $\mathbb{N}$ into subsequences $\{\mathbb{N}'\}$ as above. Along each $\mathbb{N}'$,

$$\begin{aligned}
&r_n\big(\hat\rho^*_{n,P_n} - \rho_{P_n}\big) \rightsquigarrow Z'' \in \mathbb{D}_0 \quad \text{in} \quad \mathbb{D},\\
&r_n(\hat\rho_{n,P_n} - \rho_{P_n}) \rightsquigarrow Z' \in \mathbb{D}_0 \quad \text{in} \quad \mathbb{D},\\
&\rho_{P_n} \to \rho' \quad \text{in} \quad \mathbb{D}_\rho \ \big(n \in \mathbb{N}'\big),
\end{aligned}$$

where Z'' is a separable process in $\mathbb{D}_0$ (which is given by Z' plus its independent copy $\bar{Z}'$). Indeed, note that $r_n(\hat{\rho}^*_{\rho_{n,P_n}} - \rho_{P_n}) = Z^*_{n,P_n} + Z_{n,P_n}$, and (Z^*_{n,P_n}, Z_{n,P_n}) converge weakly unconditionally to $(\bar{Z}', Z')$ by a standard argument.

Given each $\mathbb{N}'$, the proof is similar to the proof of Theorem 3.9.15 of van der Vaart and Wellner (1996). We can assume without loss of generality that the derivative $\phi'_{\rho'} : \mathbb{D} \to \mathbb{E}$ is defined and continuous on the whole of $\mathbb{D}$. Otherwise, if $\phi'_{\rho'}$ is defined and continuous only on $\mathbb{D}_0$, we can extend it to $\mathbb{D}$ by a Hahn–Banach extension such that $C = \|\phi'_{\rho'}\|_{\mathbb{D}_0 \to \mathbb{E}} = \|\phi'_{\rho'}\|_{\mathbb{D} \to \mathbb{E}} < \infty$; see van der Vaart and Wellner (1996, p. 380) for details. For each $\mathbb{N}'$, by claim (B.5), applied to $\hat{\rho}_{n,P_n}$ and to $\hat{\rho}^*_{n,P_n}$ replacing $\hat{\rho}_{n,P_n}$,

$$r_n\big(\phi(\hat{\rho}_{n,P_n}) - \phi(\rho_{P_n})\big) = \phi'_{\rho'}\big(r_n(\hat{\rho}_{n,P_n} - \rho_{P_n})\big) + o^*_{P_n}(1),$$
$$r_n\big(\phi\big(\hat{\rho}^*_{n,P_n}\big) - \phi(\rho_{P_n})\big) = \phi'_{\rho'}\big(r_n\big(\hat{\rho}^*_{n,P_n} - \rho_{P_n}\big)\big) + o^*_{P_n}(1).$$

Subtracting these equations, conclude that for each $\varepsilon > 0$,

$$\text{(B.6)} \qquad \mathrm{E}_{P_n} 1\big(\big\| r_n\big(\phi\big(\hat{\rho}^*_{n,P_n}\big) - \phi(\hat{\rho}_{n,P_n})\big) - \phi'_{\rho'}\big(r_n\big(\hat{\rho}^*_{n,P_n} - \hat{\rho}_{n,P_n}\big)\big)\big\|^*_{\mathbb{E}} > \varepsilon\big) \to 0 \quad \big(n \in \mathbb{N}'\big).$$

For every $h \in \mathrm{BL}_1(\mathbb{E})$, the function $h \circ \phi'_{\rho'}$ is contained in $\mathrm{BL}_C(\mathbb{D})$. Moreover, $r_n(\hat{\rho}^*_{n,P} - \hat{\rho}_{n,P}) \rightsquigarrow_B Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$ implies $r_n(\hat{\rho}^*_{n,P} - \hat{\rho}_{n,P}) \rightsquigarrow_B Z'$ along the subsequence $n \in \mathbb{N}'$. These two facts imply that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \big| \mathrm{E}_{B_n} h\big(\phi'_{\rho'}\big(r_n\big(\hat{\rho}^*_{n,P_n} - \hat{\rho}_{n,P_n}\big)\big)\big) - \mathrm{E}h\big(\phi_{\rho'}(Z')\big)\big| = o^*_{P_n}(1) \quad \big(n \in \mathbb{N}'\big).$$

Next, for each $\varepsilon > 0$ and along $n \in \mathbb{N}'$,

$$\begin{aligned}
&\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \big| \mathrm{E}_{B_n} h\big(r_n\big(\phi\big(\hat{\rho}^*_{n,P_n}\big) - \phi(\hat{\rho}_{n,P_n})\big)\big) - \mathrm{E}_{B_n} h\big(\phi'_{\rho'}\big(r_n\big(\hat{\rho}^*_{n,P_n} - \hat{\rho}_{n,P_n}\big)\big)\big)\big| \\
&\quad \le \varepsilon + 2\mathrm{E}_{B_n} 1\big(\big\| r_n\big(\phi\big(\hat{\rho}^*_{n,P_n}\big) - \phi(\hat{\rho}_{n,P_n})\big) \\
&\qquad - \phi'_{\rho'}\big(r_n\big(\hat{\rho}^*_{n,P_n} - \hat{\rho}_{n,P_n}\big)\big)\big\|^*_{\mathbb{E}} > \varepsilon\big) \\
&\quad = o_{P_n}(1),
\end{aligned}$$

where the $o_{P_n}(1)$ conclusion follows by the Markov inequality and by (B.6). Conclude that

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \big| \mathrm{E}_{B_n} h\big(r_n\big(\phi\big(\hat{\rho}^*_{n,P_n}\big) - \phi(\hat{\rho}_{n,P_n})\big)\big) - \mathrm{E}h\big(\phi_{\rho'}(Z')\big)\big| = o^*_{P_n}(1) \quad \big(n \in \mathbb{N}'\big).$$

Q.E.D.

APPENDIX C: KEY TOOLS II: PROBABILISTIC INEQUALITIES

Let $(W_i)_{i=1}^n$ be a sequence of i.i.d. copies of random element W taking values in the measure space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ according to probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f : \mathcal{W} \longmapsto \mathbb{R}$, equipped with a measurable envelope $F : \mathcal{W} \longmapsto \mathbb{R}$.

The following maximal inequality is due to Chernozhukov, Chetverikov, and Kato (2014).

LEMMA C.1—A Maximal Inequality: *Work with the setup above. Suppose that* $F \geq \sup_{f\in\mathcal{F}}|f|$ *is a measurable envelope with* $\|F\|_{P,q} < \infty$ *for some* $q \geq 2$. *Let* $M = \max_{i\leq n} F(W_i)$ *and* $\sigma^2 > 0$ *be any positive constant such that* $\sup_{f\in\mathcal{F}} \|f\|_{P,2}^2 \leq \sigma^2 \leq \|F\|_{P,2}^2$. *Suppose that there exist constants* $a \geq e$ *and* $v \geq 1$ *such that* $\log\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leq v(\log a + \log(1/\epsilon))$, $0 < \epsilon \leq 1$. *Then*

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

where K *is an absolute constant. Moreover, for every* $t \geq 1$, *with probability* $> 1 - t^{-q/2}$,

$$\begin{aligned}\|\mathbb{G}_n\|_{\mathcal{F}} &\leq (1+\alpha)\mathrm{E}_P\big[\|\mathbb{G}_n\|_{\mathcal{F}}\big] \\ &\quad + K(q)\big[\big(\sigma + n^{-1/2}\|M\|_{\mathrm{P}_P,q}\big)\sqrt{t} + \alpha^{-1}n^{-1/2}\|M\|_{\mathrm{P}_P,2}t\big], \\ &\quad \forall\alpha > 0,\end{aligned}$$

where $K(q) > 0$ *is a constant depending only on* q. *In particular, setting* $a \geq n$ *and* $t = \log n$, *with probability* $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leq K(q,c)\left(\sigma\sqrt{v\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P}_P,q}}{\sqrt{n}}\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)\right), \tag{C.1}$$

where $\|M\|_{\mathrm{P}_P,q} \leq n^{1/q}\|F\|_{P,q}$ *and* $K(q,c) > 0$ *is a constant depending only on* q *and* c.

APPENDIX D: PROOFS FOR SECTION 4

These results follow from the application of results given in Section 5. The details are given in the Supplemental Material.

APPENDIX E: PROOFS FOR SECTION 5

E.1. *Proof of Theorem 5.1*

In the proof $a \lesssim b$ means that $a \leq Ab$, where the constant A depends on the constants in Assumptions 5.1–5.3, but not on n once $n \geq n_0$, and not on $P \in \mathcal{P}_n$. Since the argument is asymptotic, we can assume that $n \geq n_0$ in what follows. In order to establish the result uniformly in $P \in \mathcal{P}_n$, it suffices to establish the result under the probability measure induced by any sequence $P = P_n \in \mathcal{P}_n$. In the proof we shall use P, suppressing the dependency of P_n on the sample size n. Also, let

$$B(W) := \max_{j\in[d_\theta],k\in[d_\theta+d_t]}\ \sup_{\nu\in\Theta_u\times T_u(Z_u),u\in\mathcal{U}}\big|\partial_{\nu_k}\mathrm{E}_P\big[\psi_{uj}(W_u,\nu)|Z_u\big]\big|. \tag{E.1}$$

Step 1 (A Preliminary Rate Result). In this step we claim that w.p. $1 - o(1)$, $\sup_{u\in\mathcal{U}}\|\hat{\theta}_u - \theta_u\| \lesssim \tau_n$. By definition,

$$\begin{aligned}&\big\|\mathbb{E}_n\psi_u\big(W_u,\hat{\theta}_u,\hat{h}_u(Z_u)\big)\big\| \\ &\quad \leq \inf_{\theta\in\Theta_u}\big\|\mathbb{E}_n\psi_u\big(W_u,\theta,\hat{h}_u(Z_u)\big)\big\| + \epsilon_n \quad \text{for each } u \in \mathcal{U},\end{aligned}$$

which implies via the triangle inequality that, uniformly in $u \in \mathcal{U}$ with probability $1 - o(1)$,

$$\text{(E.2)} \qquad \left\| P\big[\psi_u\big(W_u, \hat{\theta}_u, h_u(Z_u)\big)\big]\right\| \le \epsilon_n + 2\mathrm{I}_1 + 2\mathrm{I}_2 \lesssim \tau_n,$$

for I_1 and I_2 defined in Step 2 below. The $\lesssim$ bound in (E.2) follows from Step 2 and from the assumption $\epsilon_n = o(n^{-1/2})$. Since by Assumption 5.1(iv), $2^{-1}(\|J_u(\hat{\theta}_u - \theta_u)\| \wedge c_0)$ does not exceed the left side of (E.2) and $\inf_{u\in\mathcal{U}} \operatorname{mineig}(J_u' J_u)$ is bounded away from zero uniformly in n, we conclude that $\sup_{u\in\mathcal{U}} \|\hat{\theta}_u - \theta_u\| \lesssim (\inf_{u\in\mathcal{U}} \operatorname{mineig}(J_u'J_u))^{-1/2}\tau_n \lesssim \tau_n$.

Step 2 (Define and Bound I_1 and I_2). We claim that with probability $1 - o(1)$,

$$\mathrm{I}_1 := \sup_{\theta\in\Theta_u, u\in\mathcal{U}} \left\| \mathbb{E}_n \psi_u\big(W_u, \theta, \hat{h}_u(Z_u)\big) - \mathbb{E}_n \psi_u\big(W_u, \theta, h_u(Z_u)\big)\right\| \lesssim \tau_n,$$
$$\mathrm{I}_2 := \sup_{\theta\in\Theta_u, u\in\mathcal{U}} \left\| \mathbb{E}_n \psi_u\big(W_u, \theta, h_u(Z_u)\big) - P\psi_u\big(W_u, \theta, h_u(Z_u)\big)\right\| \lesssim \tau_n.$$

To establish this, we can bound $\mathrm{I}_1 \le 2\mathrm{I}_{1a} + \mathrm{I}_{1b}$ and $\mathrm{I}_2 \le \mathrm{I}_{1a}$, where with probability $1 - o(1)$,

$$\begin{aligned}
\mathrm{I}_{1a} &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}\cup\{h_u\}} \left\| \mathbb{E}_n \psi_u\big(W_u, \theta, h(Z_u)\big) - P\psi_u\big(W_u, \theta, h(Z_u)\big)\right\| \\
&\lesssim \tau_n, \\
\mathrm{I}_{1b} &:= \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}\cup\{h_u\}} \left\| P\psi_u\big(W_u, \theta, h(Z_u)\big) - P\psi_u\big(W_u, \theta, h_u(Z_u)\big)\right\| \\
&\lesssim \tau_n.
\end{aligned}$$

These bounds in turn hold by the following arguments. In order to bound I_{1b}, we employ Taylor's expansion and the triangle inequality. For $\bar{h}(Z, u, j, \theta)$ denoting a point on a line connecting vectors $h(Z_u)$ and $h_u(Z_u)$, and t_m denoting the mth element of the vector t,

$$\begin{aligned}
\mathrm{I}_{1b} &\le \sum_{j=1}^{d_\theta}\sum_{m=1}^{d_t} \sup_{\theta\in\Theta_u, u\in\mathcal{U}, h\in\mathcal{H}_{un}} \Big| P\big[\partial_{t_m} P\big[\psi_{uj}\big(W_u, \theta, \bar{h}(Z, u, j, \theta)\big) | Z_u\big] \\
&\qquad \times \big(h_m(Z_u) - h_{um}(Z_u)\big)\big]\Big| \\
&\le d_\theta d_t \|B\|_{P,2} \max_{u\in\mathcal{U}, h\in\mathcal{H}_{un}, m\in[d_t]} \|h_m - h_{um}\|_{P,2},
\end{aligned}$$

where the last inequality holds by the definition of $B(W)$ given earlier and Hölder's inequality. By Assumption 5.2(ii)(c), $\|B\|_{P,2} \le C$, and by Assumption 5.3, $\sup_{u\in\mathcal{U}, h\in\mathcal{H}_{un}, m\in[d_t]} \|h_m - h_{um}\|_{P,2} \lesssim \tau_n$, hence we conclude that $\mathrm{I}_{1b} \lesssim \tau_n$ since d_θ and d_t are fixed.

In order to bound I_{1a}, we employ the maximal inequality of Lemma C.1 to the class

$$\mathcal{F}_1 = \big\{\psi_{uj}\big(W_u, \theta, h(Z_u)\big) : j \in [d_\theta], u \in \mathcal{U}, \theta \in \Theta_u, h \in \mathcal{H}_{un} \cup \{h_u\}\big\},$$

defined in Assumption 5.3 and equipped with an envelope $F_1 \le F_0$, to conclude that with probability $1 - o(1)$,

$$\mathrm{I}_{1a} \lesssim n^{-1/2}\big(\sqrt{s_n \log(a_n)} + n^{-1/2} s_n n^{1/q} \log(a_n)\big) \lesssim \tau_n.$$

Here we use that $\log\sup_Q N(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leq s_n \log(a_n/\epsilon) \vee 0$ by Assumption 5.3; $\|F_0\|_{P,q} \leq C$ and $\sup_{f\in\mathcal{F}_1}\|f\|^2_{P,2} \leq \sigma^2 \leq \|F_0\|^2_{P,2}$ for $c \leq \sigma \leq C$ by Assumption 5.2(i); $a_n \geq n$ and $s_n \geq 1$ by Assumption 5.3.

Step 3 (Linearization). By definition,

$$\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u, \hat\theta_u, \hat h_u(Z_u)\big)\big\| \leq \inf_{\theta\in\Theta_u}\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u, \theta, \hat h_u(Z_u)\big)\big\| + \epsilon_n n^{1/2}.$$

Application of Taylor's theorem gives that, for all $u\in\mathcal{U}$,

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_n\psi_u\big(W_u, \hat\theta_u, \hat h_u(Z_u)\big) \\
&\quad = \sqrt{n}\mathbb{E}_n\psi_u\big(W_u, \theta_u, h_u(Z_u)\big) + J_u\sqrt{n}(\hat\theta_u - \theta_u) \\
&\qquad + \mathrm{D}_{u,0}(\hat h_u - h_u) + \mathrm{II}_1(u) + \mathrm{II}_2(u),
\end{aligned}$$

where the terms $\mathrm{II}_1(u)$ and $\mathrm{II}_2(u)$ are defined in Step 4 and $\mathrm{D}_{u,0}(\hat h_u - h_u)$ is treated in the next paragraph. Then, by the triangle inequality for all $u\in\mathcal{U}$ and Steps 4 and 5, we have

$$\begin{aligned}
&\big\|\sqrt{n}\mathbb{E}_n\psi_u\big(W_u, \theta_u, h_u(Z_u)\big) + J_u\sqrt{n}(\hat\theta_u - \theta_u) + \mathrm{D}_{u,0}(\hat h_u - h_u)\big\| \\
&\quad \leq \epsilon_n\sqrt{n} + \sup_{u\in\mathcal{U}}\Big(\inf_{\theta\in\Theta_u}\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u, \theta, \hat h_u(Z_u)\big)\big\| \\
&\qquad + \big\|\mathrm{II}_1(u)\big\| + \big\|\mathrm{II}_2(u)\big\|\Big) \\
&\quad = o_P(1),
\end{aligned}$$

where the $o_P(1)$ bound follows from Step 4, $\epsilon_n\sqrt{n} = o(1)$ by assumption, and Step 5.

Moreover, by the orthogonality condition,

$$\begin{aligned}
\mathrm{D}_{u,0}(\hat h_u - h_u) &:= \Bigg(\sum_{m=1}^{d_t}\sqrt{n}P\big[\partial_{t_m}P\big[\psi_{uj}\big(W_u, \theta_u, h_u(Z_u)\big)|Z_u\big] \\
&\qquad \times \big(\hat h_m(Z_u) - h_{um}(Z_u)\big)\big]\Bigg)_{j=1}^{d_\theta} \\
&= 0.
\end{aligned}$$

Conclude using Assumption 5.1(iv) that

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\big\|J_u^{-1}\sqrt{n}\mathbb{E}_n\psi_u\big(W_u, \theta_u, h_u(Z_u)\big) + \sqrt{n}(\hat\theta_u - \theta_u)\big\| \\
&\quad \leq o_P(1)\sup_{u\in\mathcal{U}}\big(\mathrm{mineig}\big(J_u'J_u\big)^{-1/2}\big) \\
&\quad = o_P(1).
\end{aligned}$$

Furthermore, the empirical process $(-\sqrt{n}\mathbb{E}_n J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u)))_{u\in\mathcal{U}}$ is equivalent to an empirical process $\mathbb{G}_n$ indexed by $\mathcal{F}_P := \{\bar\psi_{uj} : j \in [d_\theta], u \in \mathcal{U}\}$, where $\bar\psi_{uj}$ is the jth element of $-J_u^{-1}\psi_u(W_u, \theta_u, h_u(Z_u))$ and we make explicit the dependence of $\mathcal{F}_P$ on P. Let $\mathcal{M} = \{M_{ujk} : j, k \in [d_\theta], u \in \mathcal{U}\}$, where M_{ujk} is the (j, k) element of the matrix J_u^{-1}. $\mathcal{M}$ is a class of uniformly Hölder continuous functions on $(\mathcal{U}, d_{\mathcal{U}})$ with a uniform covering entropy bounded by $C \log(e/\epsilon) \vee 0$ and equipped with a constant envelope C, given the stated assumptions. This result follows from the fact that, by Assumption 5.2(ii)(b),

$$\text{(E.3)} \qquad \max_{j,k\in[d_\theta]} |M_{ujk} - M_{\bar{u}jk}| \leq \|J_u^{-1} - J_{\bar{u}}^{-1}\| = \|J_u^{-1}(J_u - J_{\bar{u}})J_{\bar{u}}^{-1}\|$$

$$\leq \|J_u - J_{\bar{u}}\| \sup_{\tilde{u}\in\mathcal{U}} \|J_{\tilde{u}}^{-1}\|^2 \lesssim \|u - \bar{u}\|^{\alpha_2},$$

and the constant envelope follows by Assumption 5.1(iv). Since $\mathcal{F}_P$ is generated as a finite sum of products of the elements of $\mathcal{M}$ and the class $\mathcal{F}_0$ defined in Assumption 5.2, the properties of $\mathcal{M}$ and the conditions on $\mathcal{F}_0$ in Assumption 5.2(ii) imply that $\mathcal{F}_P$ has a uniformly well-behaved uniform covering entropy by Lemma L.1 , namely,

$$\sup_{P\in\mathcal{P}=\bigcup_{n\geq n_0}\mathcal{P}_n} \log \sup_Q N(\epsilon\|CF_0\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2}) \lesssim \log(e/\epsilon) \vee 0,$$

where $F_P = CF_0$ is an envelope for $\mathcal{F}_P$ since

$$\sup_{f\in\mathcal{F}_P} |f| \lesssim \sup_{u\in\mathcal{U}} \|J_u^{-1}\| \sup_{f\in\mathcal{F}_0} |f| \leq CF_0$$

by Assumption 5.2(i). The class $\mathcal{F}_P$ is therefore Donsker uniformly in P because $\sup_{P\in\mathcal{P}} \|F_P\|_{P,q} \leq C \sup_{P\in\mathcal{P}} \|F_0\|_{P,q}$ is bounded by Assumption 5.2(ii), and $\sup_{P\in\mathcal{P}} \|\bar{\psi}_u - \bar{\psi}_{\bar{u}}\|_{P,2} \to 0$ as $d_{\mathcal{U}}(u, \bar{u}) \to 0$ by Assumption 5.2(b) and (E.3). Application of Theorem B.1 gives the results of the theorem.

Step 4 (Define and Bound $\mathrm{II}_1(u)$ and $\mathrm{II}_2(u)$). Let $\mathrm{II}_1(u) := (\mathrm{II}_{1j}(u))_{j=1}^{d_\theta}$ and $\mathrm{II}_2(u) = (\mathrm{II}_{2j}(u))_{j=1}^{d_\theta}$, where

$$\mathrm{II}_{1j}(u) := \sum_{r,k=1}^{d_\nu} \sqrt{n} P\big[\partial_{\nu_k}\partial_{\nu_r} P\big[\psi_{uj}\big(W_u, \bar{\nu}_u(Z_u, j)\big)|Z_u\big]$$

$$\times \big\{\hat{\nu}_{ur}(Z_u) - \nu_{ur}(Z_u)\big\}\big\{\hat{\nu}_{uk}(Z_u) - \nu_{uk}(Z_u)\big\}\big],$$

$$\mathrm{II}_{2j}(u) := \mathbb{G}_n\big(\psi_{uj}\big(W_u, \hat{\theta}_u, \hat{h}_u(Z_u)\big) - \psi_{uj}\big(W_u, \theta_u, h_u(Z_u)\big)\big),$$

$\nu_u(Z_u) := (\nu_{uk}(Z_u))_{k=1}^{d_\nu} := (\theta_u', h_u(Z_u)')'$, $\hat{\nu}_u(Z_u) := (\hat{\nu}_{uk}(Z_u))_{k=1}^{d_\nu} := (\hat{\theta}_u', \hat{h}_u(Z_u)')'$, $d_\nu = d_\theta + d_t$, and $\bar{\nu}_u(Z_u, j)$ is a vector on the line connecting $\nu_u(Z_u)$ and $\hat{\nu}_u(Z_u)$.

First, by Assumptions 5.2(ii)(d) and 5.3, the claim of Step 1, and the Hölder inequality,

$$\max_{j\in[d_\theta]} \sup_{u\in\mathcal{U}} |\mathrm{II}_{1j}(u)|$$

$$\leq \sup_{u\in\mathcal{U}} \sum_{r,k=1}^{d_\nu} \sqrt{n} P\big[C|\hat{\nu}_{ur}(Z_u) - \nu_{ur}(Z_u)||\hat{\nu}_{uk}(Z_u) - \nu_{uk}(Z_u)|\big]$$

$$\leq C\sqrt{n} d_\nu^2 \max_{k\in[d_\nu]} \sup_{u\in\mathcal{U}} \|\hat{\nu}_{uk} - \nu_{uk}\|_{P,2}^2$$

$$\lesssim_P \sqrt{n}\tau_n^2 = o(1).$$

Second, we have that with probability $1 - o(1)$, $\max_{j\in[d_\theta]}\sup_{u\in\mathcal{U}}|\mathrm{II}_{2j}(u)| \lesssim \sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|$, where, for $\Theta_{un} := \{\theta\in\Theta_u : \|\theta-\theta_u\|\le C\tau_n\}$,

$$\begin{aligned}\mathcal{F}_2 = \big\{&\psi_{uj}\big(W_u,\theta,h(Z_u)\big)-\psi_{uj}\big(W_u,\theta_u,h_u(Z_u)\big):\\ &j\in[d_\theta],u\in\mathcal{U},h\in\mathcal{H}_{un},\theta\in\Theta_{un}\big\}.\end{aligned}$$

Application of Lemma C.1 with an envelope $F_2\lesssim F_0$ gives that, with probability $1-o(1)$,

$$\sup_{f\in\mathcal{F}_2}\big|\mathbb{G}_n(f)\big|\lesssim \tau_n^{\alpha/2}\sqrt{s_n\log(a_n)}+n^{-1/2}s_n n^{1/q}\log(a_n), \tag{E.4}$$

since $\sup_{f\in\mathcal{F}_2}|f|\le 2\sup_{f\in\mathcal{F}_1}|f|\le 2F_0$ by Assumption 5.3; $\|F_0\|_{P,q}\le C$ by Assumption 5.2(i); $\log\sup_Q N(\epsilon\|F_2\|_{Q,2},\mathcal{F}_2,\|\cdot\|_{Q,2})\lesssim (s_n\log a_n + s_n\log(a_n/\epsilon))\vee 0$ by Lemma K.1 because $\mathcal{F}_2=\mathcal{F}_1-\mathcal{F}_0$ for the $\mathcal{F}_0$ and $\mathcal{F}_1$ defined in Assumptions 5.2(i) and 5.3; and σ can be chosen so that $\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}\le\sigma\lesssim\tau_n^{\alpha/2}$. Indeed,

$$\begin{aligned}&\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}^2\\ &\quad\le \sup_{j\in[d_\theta],u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} P\big(P\big[\big(\psi_{uj}\big(W_u,\nu(Z_u)\big)-\psi_{uj}\big(W_u,\nu_u(Z_u)\big)\big)^2|Z_u\big]\big)\\ &\quad\le \sup_{u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} P\big(C\big\|\nu(Z_u)-\nu_u(Z_u)\big\|^\alpha\big)\\ &\quad= \sup_{u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} C\|\nu-\nu_u\|_{P,\alpha}^\alpha\le \sup_{u\in\mathcal{U},\nu\in\Theta_{un}\times\mathcal{H}_{un}} C\|\nu-\nu_u\|_{P,2}^\alpha\lesssim\tau_n^\alpha,\end{aligned}$$

where the first inequality follows by the law of iterated expectations; the second inequality follows by Assumption 5.2(ii)(a); and the last inequality follows from $\alpha\in[1,2]$ by Assumption 5.2, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha\in[1,\infty]$, and Assumption 5.3.

Conclude using the growth conditions of Assumption 5.3 that, with probability $1-o(1)$,

$$\max_{j\in[d_\theta]}\sup_{u\in\mathcal{U}}\big|\mathrm{II}_{2j}(u)\big|\lesssim\tau_n^{\alpha/2}\sqrt{s_n\log(a_n)}+n^{-1/2}s_n n^{1/q}\log(a_n)=o(1). \tag{E.5}$$

Step 5. In this step we show that $\sup_{u\in\mathcal{U}}\inf_{\theta\in\Theta_u}\sqrt{n}\|\mathbb{E}_n\psi_u(W_u,\theta,\hat{h}_u(Z_u))\|=o_P(1)$. We have that, with probability $1-o(1)$,

$$\inf_{\theta\in\Theta_u}\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u,\theta,\hat{h}_u(Z_u)\big)\big\|\le\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u,\bar\theta_u,\hat{h}_u(Z_u)\big)\big\|,$$

where $\bar\theta_u=\theta_u-J_u^{-1}\mathbb{E}_n\psi_u(W_u,\theta_u,h_u(Z_u))$, since $\bar\theta_u\in\Theta_u$ for all $u\in\mathcal{U}$ with probability $1-o(1)$, and, in fact, $\sup_{u\in\mathcal{U}}\|\bar\theta_u-\theta_u\|=O_P(1/\sqrt{n})$ by the last paragraph of Step 3.

Then, arguing similarly to Steps 3 and 4, we can conclude that uniformly in $u\in\mathcal{U}$,

$$\begin{aligned}&\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u,\bar\theta_u,\hat{h}_u(Z_u)\big)\big\|\\ &\quad\le\sqrt{n}\big\|\mathbb{E}_n\psi_u\big(W_u,\theta_u,h_u(Z_u)\big)+J_u(\bar\theta_u-\theta_u)+\mathrm{D}_{u,0}(\hat{h}_u-h_u)\big\|\\ &\qquad+o_P(1),\end{aligned}$$

where the first term on the right side is zero by definition of $\bar\theta_u$ and $\mathrm{D}_{u,0}(\hat{h}_u-h_u)=0$. *Q.E.D.*

E.2. *Proof of Theorem 5.2*

Step 0. In the proof, $a \lesssim b$ means that $a \leq Ab$, where the constant A depends on the constants in Assumptions 5.1– 5.3, but not on n once $n \geq n_0$, and not on $P \in \mathcal{P}_n$. In Step 1, we consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, suppressing the index n. Since the argument is asymptotic, we can assume that $n \geq n_0$ in what follows.

Let $\mathbb{P}_n$ denote the measure that puts mass n^{-1} at the points (ξ_i, W_i) for $i = 1, \dots, n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_n f = n^{-1} \sum_{i=1}^n f(\xi_i, W_i)$, and $\mathbb{G}_n$ denote the corresponding empirical process $\sqrt{n}(\mathbb{E}_n - P)$, that is,

$$\begin{aligned} \mathbb{G}_n f &= \sqrt{n}(\mathbb{E}_n f - Pf) \\ &= n^{-1/2} \sum_{i=1}^n \left(f(\xi_i, W_i) - \int f(s, w) \, dP_\xi(s) \, dP(w) \right). \end{aligned}$$

Recall that we define the bootstrap draw as

$$Z^*_{n,P} := \sqrt{n}\big(\hat{\theta}^* - \hat{\theta}\big) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \hat{\psi}_u(W_i) \right)_{u \in \mathcal{U}} = (\mathbb{G}_n \xi \hat{\psi}_u)_{u \in \mathcal{U}},$$

where $\hat{\psi}_u(W) = -\hat{J}_u^{-1} \psi_u(W_u, \hat{\theta}_u, \hat{h}_u(Z_u))$.

Step 1 (Linearization). In this step we establish that

$$\zeta^*_{n,P} := Z^*_{n,P} - G^*_{n,P} = o_P(1) \quad \text{in} \quad \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}, \tag{E.6}$$

where $G^*_{n,P} := (\mathbb{G}_n \xi \bar{\psi}_u)_{u \in \mathcal{U}}$, and $\bar{\psi}_u(W) = -J_u^{-1} \psi_u(W_u, \theta_u, h_u(Z_u))$.

The claim would follow from demonstrating that (a)

$$Z^\star_{n,P} - G^*_{n,P} = o_P(1) \quad \text{in} \quad \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}, \tag{E.7}$$

where $Z^\star_{n,P} := (\mathbb{G}_n \xi \check{\psi}_u)_{u \in \mathcal{U}}$, and $\check{\psi}_u(W) = -J_u^{-1} \psi_u(W_u, \hat{\theta}_u, \hat{h}_u(Z_u))$ (note that the hat from J_u disappeared); and (b)

$$Z^\star_{n,P} - Z^*_{n,P} = o_P(1) \quad \text{in} \quad \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}. \tag{E.8}$$

To show claim (E.7), we note that with probability $1 - \delta_n$, $\hat{h}_u \in \mathcal{H}_{un}$, $\hat{\theta}_u \in \Theta_{un} = \{\theta \in \Theta_u : \|\theta - \theta_u\| \leq C\tau_n\}$, so that $\|\zeta^\star_{n,P}\|_{\mathbb{D}} \lesssim \sup_{f \in \mathcal{F}_3} |\mathbb{G}_n[\xi f]|$, where

$$\mathcal{F}_3 = \big\{ \tilde{\psi}_{uj}(\bar{\theta}_u, \bar{h}_u) - \bar{\psi}_{uj} : j \in [d_\theta], u \in \mathcal{U}, \bar{\theta}_u \in \Theta_{un}, \bar{h}_u \in \mathcal{H}_{un} \big\},$$

where $\tilde{\psi}_{uj}(\bar{\theta}_u, \bar{h}_u)$ is the jth element of $-J_u^{-1} \psi_u(W_u, \bar{\theta}_u, \bar{h}_u(Z_u))$, and $\bar{\psi}_{uj}$ is the jth element of $-J_u^{-1} \psi_u(W_u, \theta_u, h_u(Z_u))$. By arguments similar to those employed in the proof of the previous theorem, $\mathcal{F}_3$ obeys

$$\log \sup_Q N\big(\epsilon \|F_3\|_{Q,2}, \mathcal{F}_3, \|\cdot\|_{Q,2}\big) \lesssim \big(s_n \log a_n + s_n \log(a_n/\epsilon)\big) \vee 0,$$

for an envelope $F_3 \lesssim F_0$. By Lemma K.1, multiplication of this class by ξ does not change the entropy bound modulo an absolute constant, namely,

$$\log \sup_Q N\big(\epsilon \big\| |\xi| F_3 \big\|_{Q,2}, \xi \mathcal{F}_3, \|\cdot\|_{Q,2}\big) \lesssim \big(s_n \log a_n + s_n \log(a_n/\epsilon)\big) \vee 0.$$

Also $\mathrm{E}[\exp(|\xi|)] < \infty$ implies $(\mathrm{E}[\max_{i\le n} |\xi_i|^2])^{1/2} \lesssim \log n$, so that, using independence of $(\xi_i)_{i=1}^n$ from $(W_i)_{i=1}^n$ and Assumption 5.2(i),

$$\Big\| \max_{i\le n} \xi_i F_0(W_i) \Big\|_{\mathrm{P}_P,2} \le \Big\| \max_{i\le n} \xi_i \Big\|_{\mathrm{P}_P,2} \Big\| \max_{i\le n} F_0(W_i) \Big\|_{\mathrm{P}_P,2} \lesssim n^{1/q} \log n.$$

Applying Lemma C.1,

$$\sup_{f\in\xi\mathcal{F}_3} \big|\mathbb{G}_n(f)\big| = O_P\bigg(\tau_n^{\alpha/2} \sqrt{s_n \log(a_n)} + \frac{s_n n^{1/q} \log n}{\sqrt{n}} \log(a_n) \bigg) = o_P(1),$$

for $\sup_{f\in\xi\mathcal{F}_3} \|f\|_{P,2} = \sup_{f\in\mathcal{F}_3} \|f\|_{P,2} \lesssim \sigma_n \lesssim \tau_n^{\alpha/2}$, where the details of calculations are similar to those in the proof of Theorem 5.1. Indeed, with probability $1 - o(\delta_n)$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_3} \|f\|_{P,2}^2 \\
&\quad \lesssim \sup_{u\in\mathcal{U}} \big\| J_u^{-1} \big\|^2 \sup_{j\in[d_\theta], u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} P\Big(P\Big[\big(\psi_{uj}\big(W_u, \nu(Z_u)\big) \\
&\qquad - \psi_{uj}\big(W_u, \nu_u(Z_u)\big)\big)^2 | Z_u\Big]\Big) \\
&\quad \lesssim \sup_{u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} \|\nu - \nu_u\|_{P,\alpha}^\alpha \lesssim \sup_{u\in\mathcal{U}, \nu\in\Theta_{un}\times\mathcal{H}_{un}} \|\nu - \nu_u\|_{P,2}^\alpha \lesssim \tau_n^\alpha,
\end{aligned}$$

where the first inequality follows from the triangle inequality and the law of iterated expectations; the second inequality follows by Assumption 5.2(ii)(a), Assumption 5.2(i); the third inequality follows from $\alpha \in [1, 2]$ by Assumption 5.2, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha \in [1, \infty]$, and Assumption 5.3; and the last inequality follows from $\|\nu - \nu_u\|_{P,2} \lesssim \tau_n$ by the definition of Θ_{un} and $\mathcal{H}_{un}$. The claim (E.7) follows.

To show claim (E.8), bound

$$\begin{aligned}
\big\| Z_{n,P}^{\star} - Z_{n,P}^{*} \big\|_{\mathbb{D}} &= \big\| \big(\hat{J}_u^{-1} J_u Z_{n,P}^*(u) - Z_{n,P}^*(u) \big)_{u\in\mathcal{U}} \big\|_{\mathbb{D}} \\
&\lesssim \sup_{u\in\mathcal{U}} \big\| \hat{J}_u^{-1} J_u - I \big\| \big\| Z_{n,P}^* \big\|_{\mathbb{D}} = o_P(1),
\end{aligned}$$

since $\sup_{u\in\mathcal{U}} \|\hat{J}_u^{-1} J_u - I\| = o_P(1)$ by the assumption of the theorem, and since $\|Z_{n,P}^*\|_{\mathbb{D}} = O_P(1)$ by $\|Z_{n,P}^*\|_{\mathbb{D}} = \|G_{n,P}^* + o_P(1)\|_{\mathbb{D}} \rightsquigarrow_B \|Z_P\|_{\mathbb{D}}$, which follows by claim (E.7) and by $G_{n,P}^* \rightsquigarrow_B Z_P$ in $\mathbb{D}$ holding by Theorem B.2.

Step 2. Here we are claiming that $Z_{n,P}^* \rightsquigarrow_B Z_P$ in $\mathbb{D} = \ell^\infty(\mathcal{U})^{d_\theta}$, under any sequence $P = P_n \in \mathcal{P}_n$, where $Z_P = (\mathbb{G}_P \bar{\psi}_u)_{u\in\mathcal{U}}$. By the triangle inequality and Step 1,

$$\begin{aligned}
&\sup_{h\in\mathrm{BL}_1(\mathbb{D})} \big| \mathrm{E}_{B_n} h\big(Z_{n,P}^*\big) - \mathrm{E}_P h(Z_P) \big| \\
&\quad \le \sup_{h\in\mathrm{BL}_1(\mathbb{D})} \big| \mathrm{E}_{B_n} h\big(G_{n,P}^*\big) - \mathrm{E}_P h(Z_P) \big| + \mathrm{E}_{B_n}\big(\big\| \zeta_{n,P}^* \big\|_{\mathbb{D}} \wedge 2 \big),
\end{aligned}$$

where the first term is $o_P^*(1)$, since $G_{n,P}^* \rightsquigarrow_B Z_P$ by Theorem B.2, and the second term is $o_P(1)$ because $\|\zeta_{n,P}^*\|_{\mathbb{D}} = o_P(1)$ implies that $\mathrm{E}_P(\|\zeta_{n,P}^*\|_{\mathbb{D}} \wedge 2) = \mathrm{E}_P \mathrm{E}_{B_n}(\|\zeta_{n,P}^*\|_{\mathbb{D}} \wedge 2) \to 0$, which in turn implies that $\mathrm{E}_{B_n}(\|\zeta_{n,P}^*\|_{\mathbb{D}} \wedge 2) = o_P(1)$ by the Markov inequality. *Q.E.D.*

E.3. *Proof of Theorem 5.3*

This is an immediate consequence of Theorems 5.1, 5.2, B.3, and B.4.

APPENDIX F: IMPLEMENTATION DETAILS

In this section, we provide details about how we implemented the methodology developed in the main body of the paper in the empirical example. We first discuss estimation of local average treatment effects (LATE) and then extend this discussion to estimation of local quantile treatment effects (LQTE). Estimation of all other quantities proceeds in a similar fashion and so is not discussed.

F.1. *Local Average Treatment Effects*

Recall that the LATE of treatment D on outcome Y is defined as

$$\begin{aligned}\Delta_{\text{LATE}} &= \theta_Y(1) - \theta_Y(0) \\ &= \frac{\alpha_{1_1(D)Y}(1) - \alpha_{1_1(D)Y}(0)}{\alpha_{1_1(D)}(1) - \alpha_{1_1(D)}(0)} - \frac{\alpha_{1_0(D)Y}(1) - \alpha_{1_0(D)Y}(0)}{\alpha_{1_0(D)}(1) - \alpha_{1_0(D)}(0)}\end{aligned}$$

for $\alpha_V(z)$ and $\theta_Y(d)$ defined in equations (2.1) and (2.3), respectively. It then follows by plugging in the definition of $\alpha_V(z)$ that we can express the LATE as

$$\Delta_{\text{LATE}} = \frac{\alpha_Y(1) - \alpha_Y(0)}{\alpha_{1_1(D)}(1) - \alpha_{1_1(D)}(0)}.$$

To obtain an estimate of the LATE, we thus need estimates of $\alpha_Y(z)$ and $\alpha_{1_1(D)}(z)$. Using the low-bias moment function given in equation (3.13), estimates of these key quantities can be constructed from estimates of $\mathrm{E}_P[Y|Z=1,X]$, $\mathrm{E}_P[Y|Z=0,X]$, $\mathrm{E}_P[D|Z=1,X]$, $\mathrm{E}_P[D|Z=0,X]$, and $\mathrm{E}_P[Z|X]$ where Z is the binary instrument (401(k) eligibility); D is the binary treatment (401(k) participation); X is the set of raw covariates discussed in the empirical section; and Y is net financial assets. In our application, we have $\mathrm{E}_P[D|Z=0,X]=0$ since one cannot participate unless one is eligible. We use Post-Lasso to estimate $\mathrm{E}_P[Y|Z=1,X]$ and $\mathrm{E}_P[Y|Z=0,X]$ and post-ℓ_1-penalized logistic regression to estimate $\mathrm{E}_P[D|Z=1,X]$ and $\mathrm{E}_P[Z|X]$.

To estimate $\mathrm{E}_P[Y|Z=1,X]$, we postulate that $\mathrm{E}_P[Y|Z=1,X]\approx f(X)'\beta_Y(1)$, where $f(X)$ is one of the prespecified sets of controls discussed in the empirical section with dimension p. Let $\mathcal{I}_1$ denote the indices of observations that have $z_i=1$. To estimate the coefficients $\beta_Y(1)$, we apply the formulation of the Post-Lasso estimator given in Belloni et al. (2012) with outcomes $\{y_i\}_{i\in\mathcal{I}_1}$ and covariates $\{f(x_i)\}_{i\in\mathcal{I}_1}$. We set $\lambda = 1.1\sqrt{n}\Phi^{-1}(1-(0.1/\log(n))/(2(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings according to Algorithm A.1 of Belloni et al. (2012) using Post-Lasso coefficient estimates at each iteration and with the maximum number of iterations set to 15.[31] Let $\hat{\beta}_Y(1)$ denote the resulting Post-Lasso estimates of the coefficients using λ given above and the final set of penalty loadings. We then estimate $\mathrm{E}_P[Y|Z=1,X=x_i]$ as $f(x_i)'\hat{\beta}_Y(1)$ for each $i=1,\dots,n$. We follow the same procedure

[31]Here and in all following instances, we stop iterating before reaching the maximum number of iterations if the ℓ_2-norm of the difference in penalty loadings calculated across consecutive iterations is less than 10^{-6}.

to obtain estimates of $\mathrm{E}_P[Y|Z=0, X=x_i]$ as $f(x_i)'\hat{\beta}_Y(0)$ for each $i=1,\dots,n$ where $\hat{\beta}_Y(0)$ are the Post-Lasso estimates using only the observations with $z_i=0$.

Estimation of $\mathrm{E}_P[D|Z=1, X]$ and $\mathrm{E}_P[Z|X]$ proceed similarly replacing Post-Lasso estimation with post-ℓ_1-penalized logistic regression. Specifically, we assume that $\mathrm{E}_P[D|Z=1, X]\approx \Lambda_0(f(X)'\beta_D(1))$ where $\Lambda_0(\cdot)$ is the logistic link function. We then obtain estimates of $\beta_D(1)$ by using the post-ℓ_1-penalized estimator defined in equations (3.10) and (3.11) based on the logistic link function and with outcomes $\{d_i\}_{i\in\mathcal{I}_1}$ and covariates $\{f(x_i)\}_{i\in\mathcal{I}_1}$ for $\mathcal{I}_1$ defined as above. We set $\lambda=1.1\sqrt{n}\Phi^{-1}(1-(0.1/\log(n))/(2(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 6.1 of the main text with a maximum of 15 iterations. Let $\hat{\beta}_D(1)$ denote the resulting post-ℓ_1-penalized estimates of the coefficients using λ given above and the final set of penalty loadings. We estimate $\mathrm{E}_P[D|Z=1, X=x_i]$ as $\Lambda_0(f(x_i)'\hat{\beta}_D(1))$ for each $i=1,\dots,n$. We follow this procedure to obtain estimates of $\mathrm{E}_P[Z|X=x_i]$ as $\Lambda_0(f(x_i)'\hat{\beta}_Z)$ for each $i=1,\dots,n$ where $\hat{\beta}_Z$ are the post-ℓ_1-penalized coefficient estimates obtained with $\{z_i\}_{i=1}^n$ as the outcome and $\{f(x_i)\}_{i=1}^n$ as covariates using $\lambda=1.1\sqrt{n}\Phi^{-1}(1-(0.1/\log(n))/(2p))$.

Using these baseline quantities, we obtain estimates

$$
\begin{aligned}
\hat{\alpha}_Y(1) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{z_i\big(y_i-f(x_i)'\hat{\beta}_Y(1)\big)}{\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}+f(x_i)'\hat{\beta}_Y(1)\right)\\
&= \frac{1}{n}\sum_{i=1}^n\psi_{1,i},\\
\hat{\alpha}_Y(0) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{(1-z_i)\big(y_i-f(x_i)'\hat{\beta}_Y(0)\big)}{1-\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}+f(x_i)'\hat{\beta}_Y(0)\right)\\
&= \frac{1}{n}\sum_{i=1}^n\psi_{0,i},\\
\hat{\alpha}_{1_1(D)}(1) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{z_i\big(d_i-\Lambda_0\big(f(x_i)'\hat{\beta}_D(1)\big)\big)}{\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}+\Lambda_0\big(f(x_i)'\hat{\beta}_D(1)\big)\right)\\
&= \frac{1}{n}\sum_{i=1}^n\upsilon_{1,i},\\
\hat{\alpha}_{1_1(D)}(0) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{(1-z_i)d_i}{1-\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}\right)=\frac{1}{n}\sum_{i=1}^n\upsilon_{0,i}=0.
\end{aligned}
$$

We then plug these estimates in to obtain

$$
\hat{\Delta}_{\mathrm{LATE}}=\frac{\hat{\alpha}_Y(1)-\hat{\alpha}_Y(0)}{\hat{\alpha}_{1_1(D)}(1)-\hat{\alpha}_{1_1(D)}(0)}.
$$

We report both analytic and bootstrap standard error estimates for the LATE. The analytic standard errors are calculated as

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\psi_{1,i}-\psi_{0,i}}{\hat{\alpha}_{1_1(D)}(1)-\hat{\alpha}_{1_1(D)}(0)}-\hat{\Delta}_{\text{LATE}}\right)^2 \Big/ n}.$$

We use wild bootstrap weights for obtaining the multiplier bootstrap estimates of the standard errors with 500 bootstrap replications. Specifically, for each $b = 1, \dots, 500$, we calculate a bootstrap estimate of the LATE as

$$\hat{\Delta}^b_{\text{LATE}} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(\psi_{1,i}-\psi_{0,i})\xi_i^b}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(\upsilon_{1,i}-\upsilon_{0,i})\xi_i^b},$$

where $\xi_i^b = 1 + r^b_{1,i}/\sqrt{2} + ((r^b_{2,i})^2 - 1)/2$ is the bootstrap draw for multiplier weight for observation i in bootstrap repetition b where $r^b_{1,i}$ and $r^b_{2,i}$ are random numbers generated as i.i.d. draws from two independent standard normal random variables. The bootstrap standard error estimate is then the bootstrap interquartile range rescaled with the normal distribution: $[q_{\text{LATE}}(0.75) - q_{\text{LATE}}(0.25)]/[q_N(0.75) - q_N(0.25)]$, where $q_{\text{LATE}}(p)$ is the pth quantile of $\{\hat{\Delta}^b_{\text{LATE}}\}_{b=1}^{500}$ and $q_N(p)$ is the pth quantile of the $N(0, 1)$.

F.2. *Local Quantile Treatment Effects*

Calculation and inference for LQTE is more cumbersome than for the LATE. We begin by choosing the set over which we would like to look at the LQTE. In our example, we chose to look at quantiles in the interval $[0.1, 0.9]$.

To calculate the LQTE, we first calculate the local average structural function for outcomes $Y_u = 1(Y \le u)$ for a set of u and then invert to obtain estimates of the LQTE. In our example, we chose to look at $u \in [q_Y(0.05), q_Y(0.95)]$ where $q_Y(0.05)$ and $q_Y(0.95)$ are respectively the sample 5th and 95th percentiles of the outcome of interest Y. Since looking at the continuum of values in this interval is infeasible, we discretize the interval and look at $Y_u = 1(Y \le u)$ for $u \in \{q_Y(0.05), q_Y(0.06), q_Y(0.07), \dots, q_Y(0.93), q_Y(0.94), q_Y(0.95)\}$. That is, we set u equal to each percentile of Y between the 5th and 95th percentiles for a total of 91 different values of u to be considered. For each value of u, we need an estimate of the local average structural function defined in (2.3) for $d \in \{0, 1\}$:

$$\theta_{1(Y\le u)}(d) = \frac{\alpha_{1_d(D)1(Y\le u)}(1) - \alpha_{1_d(D)1(Y\le u)}(0)}{\alpha_{1_d(D)}(1) - \alpha_{1_d(D)}(0)}.$$

As with the LATE, we need estimates of $\mathrm{E}_P[D|Z = 1, X]$ and $\mathrm{E}_P[Z|X]$. We estimate these quantities as we did for the LATE but change the value of the penalty parameter used to reflect the fact that we are now interested in a large set, in theory a continuum, of model selection problems. Specifically, we assume that $\mathrm{E}_P[D|Z = 1, X] \approx \Lambda_0(f(X)'\beta_D(1))$ where $\Lambda_0(\cdot)$ is the logistic link function and $f(X)$ is one of the prespecified sets of controls discussed in the empirical section with dimension p. We then obtain estimates of $\beta_D(1)$ by using the post-ℓ_1-penalized estimator defined in equations (3.10) and (3.11) based on the logistic link function and with outcomes $\{d_i\}_{i\in\mathcal{I}_1}$ and covariates $\{f(x_i)\}_{i\in\mathcal{I}_1}$ for $\mathcal{I}_1$ defined as above. We set $\lambda = 1.1\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2n(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 6.1 with a maximum of 15 iterations. Let $\hat{\beta}_D(1)$ denote the resulting post-ℓ_1-penalized estimates of the coefficients using λ given above and the final set of penalty loadings. We estimate $\mathrm{E}_P[D|Z=1, X=x_i]$ as $\Lambda_0(f(x_i)'\hat{\beta}_D(1))$ for each $i=1,\dots,n$. We follow this procedure to obtain estimates of $\mathrm{E}_P[Z|X]$ as $\Lambda_0(f(x_i)'\hat{\beta}_Z)$ for each $i=1,\dots,n$ where $\hat{\beta}_Z$ are the post-ℓ_1-penalized coefficient estimates obtained with $\{z_i\}_{i=1}^n$ as the outcome and $\{f(x_i)\}_{i=1}^n$ as covariates and $\lambda = 1.1\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2np))$. We also still have $\mathrm{E}_P[D|Z=0, X]=0$ in our application since one cannot participate in a 401(k) unless one is eligible. We then plug in these estimates to obtain

$$
\begin{aligned}
\hat{\alpha}_{1_1(D)}(1) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{z_i\big(d_i - \Lambda_0\big(f(x_i)'\hat{\beta}_D(1)\big)\big)}{\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)} + \Lambda_0\big(f(x_i)'\hat{\beta}_D(1)\big)\right)\\
&= \frac{1}{n}\sum_{i=1}^n \upsilon_{1,1,i},\\
\hat{\alpha}_{1_1(D)}(0) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{(1-z_i)d_i}{1-\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}\right) = \frac{1}{n}\sum_{i=1}^n \upsilon_{1,0,i} = 0,\\
\hat{\alpha}_{1_0(D)}(1) &= 1 - \hat{\alpha}_{1_1(D)}(1),\\
\hat{\alpha}_{1_0(D)}(0) &= 1 - \hat{\alpha}_{1_1(D)}(0).
\end{aligned}
$$

We also need to obtain estimates of $\alpha_{1_d(D)1(Y\le u)}(z)$ for each value of u and for

$$(z, d) \in \{(0,0), (0,1), (1,0), (1,1)\}.$$

These estimates will depend on the propensity score, $\mathrm{E}_P[Z|X]$, estimated above and quantities of the form $\mathrm{E}_P[1(D=d)1(Y\le u)|Z=z, X]$. We again approximate this function with $\mathrm{E}_P[1(D=d)1(Y\le u)|Z=z,X] \approx \Lambda_0(X'\beta_{\mathbf{1}_d(D)Y_u}(z))$ and estimate the coefficients $\beta_{\mathbf{1}_d(D)Y_u}(z)$ for each combination of d and z and each u using the post-ℓ_1-penalized estimator defined in equations (3.10) and (3.11) based on the logistic link function. We set $\lambda = 1.1\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2n(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 6.1 of the main text with a maximum of 15 iterations. We follow this procedure for each u with $\{1(y_i\le u)1(d_i=1)\}_{i\in\mathcal{I}_1}$ as the outcome and covariates $\{f(x_i)\}_{i\in\mathcal{I}_1}$, with $\{1(y_i\le u)1(d_i=0)\}_{i\in\mathcal{I}_1}$ as the outcome and covariates $\{f(x_i)\}_{i\in\mathcal{I}_1}$, and with $\{1(y_i\le u)1(d_i=0)\}_{i\in\mathcal{I}_0}$ as the outcome and covariates $\{f(x_i)\}_{i\in\mathcal{I}_0}$ for $\mathcal{I}_1$ and $\mathcal{I}_0$ defined as above to obtain point estimates $\hat{\beta}_{\mathbf{1}_1(D)Y_u}(1)$, $\hat{\beta}_{\mathbf{1}_0(D)Y_u}(1)$, and $\hat{\beta}_{\mathbf{1}_0(D)Y_u}(0)$, respectively. We then estimate $\mathrm{E}_P[1(D=1)1(Y\le u)|Z=1, X=x_i]$ as $\Lambda_0(f(x_i)'\hat{\beta}_{\mathbf{1}_1(D)Y_u}(1))$ for each $i=1,\dots,n$ and obtain estimates of $\mathrm{E}_P[1(D=0)1(Y\le u)|Z=1, X=x_i]$, and $\mathrm{E}_P[1(D=0)1(Y\le u)|Z=0, X=x_i]$ analogously. As before, we have $\mathrm{E}_P[1(D=1)1(Y\le u)|Z=0, X]=0$ since one cannot participate unless one is eligible. We then plug in these estimates to obtain

$$
\begin{aligned}
&\hat{\alpha}_{1_1(D)1(Y\le u)}(1)\\
&\quad = \frac{1}{n}\sum_{i=1}^n\left(\frac{z_i\big(d_i 1(y_i\le u) - \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_1(D)Y_u}(1)\big)\big)}{\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_1(D)Y_u}(1)\big)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \kappa_{u,1,1,i}, \\
&\hat{\alpha}_{1_1(D)1(Y\le u)}(0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-z_i)\big(d_i 1(y_i \le u)\big)}{1-\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)}\right) = \frac{1}{n}\sum_{i=1}^{n}\kappa_{u,1,0,i} = 0, \\
&\hat{\alpha}_{1_0(D)1(Y\le u)}(1) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\frac{z_i\big((1-d_i)1(y_i \le u) - \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_0(D)Y_u}(1)\big)\big)}{\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)} \\
&\quad + \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_0(D)Y_u}(1)\big)\Bigg) \\
&= \frac{1}{n}\sum_{i=1}^{n} \kappa_{u,0,1,i}, \\
&\hat{\alpha}_{1_0(D)1(Y\le u)}(0) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\frac{(1-z_i)\big((1-d_i)1(y_i \le u) - \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_0(D)Y_u}(0)\big)\big)}{1-\Lambda_0\big(f(x_i)'\hat{\beta}_Z\big)} \\
&\quad + \Lambda_0\big(f(x_i)'\hat{\beta}_{\mathbf{1}_0(D)Y_u}(0)\big)\Bigg) \\
&= \frac{1}{n}\sum_{i=1}^{n} \kappa_{u,0,0,i}.
\end{aligned}
$$

Estimates of the local average structural (distribution) functions are formed using the estimators defined in the previous two paragraphs as

$$\hat{\theta}_{1(Y\le u)}(d) = \frac{\hat{\alpha}_{1_d(D)1(Y\le u)}(1) - \hat{\alpha}_{1_d(D)1(Y\le u)}(0)}{\hat{\alpha}_{1_d(D)}(1) - \hat{\alpha}_{1_d(D)}(0)}.$$

To obtain LQTE estimates, we then need to invert these local average structural functions. Since we only have the estimated distribution for each d evaluated on the finite grid of points $u \in \{q_Y(0.05), q_Y(0.06), q_Y(0.07), \dots, q_Y(0.93), q_Y(0.94), q_Y(0.95)\}$, we do this inversion by linearly interpolating the value of the distribution function between these points to find the value of the outcome associated with each quantile in the set $q \in [0.1, 0.11, 0.12, \dots, 0.89, 0.9]$ which we denote as $\hat{\theta}_Y^{\leftarrow}(q,d)$. The LQTE at point q is then estimated as $\hat{\Delta}(q) = \hat{\theta}_Y^{\leftarrow}(q,1) - \hat{\theta}_Y^{\leftarrow}(q,0)$.

For the LQTE, we only report inference based on the multiplier bootstrap using 500 bootstrap replications. For each $b = 1, \dots, 500$, we generate bootstrap weights as $\xi_i^b = 1 + r_{1,i}^b/\sqrt{2} + ((r_{2,i}^b)^2 - 1)/2$ for observation i in bootstrap repetition b where $r_{1,i}^b$ and $r_{2,i}^b$ are random numbers generated as i.i.d. draws from two independent standard normal random variables. We then use these weights to form bootstrap estimates of the local average structural functions

$$\hat{\theta}^b_{1(Y\leq u)}(d) = \frac{\hat{\alpha}^b_{1_d(D)1(Y\leq u)}(1) - \hat{\alpha}^b_{1_d(D)1(Y\leq u)}(0)}{\hat{\alpha}^b_{1_d(D)}(1) - \hat{\alpha}^b_{1_d(D)}(0)},$$

where

$$\hat{\alpha}^b_{1_1(D)}(1) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \upsilon_{1,1,i}, \quad \hat{\alpha}^b_{1_1(D)}(0) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \upsilon_{1,0,i},$$

$$\hat{\alpha}^b_{1_0(D)}(1) = 1 - \hat{\alpha}^b_{1_1(D)}(1), \quad \hat{\alpha}^b_{1_0(D)}(0) = 1 - \hat{\alpha}^b_{1_1(D)}(0),$$

$$\hat{\alpha}^b_{1_1(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \kappa_{u,1,1,i},$$

$$\hat{\alpha}^b_{1_1(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \kappa_{u,1,0,i} = 0,$$

$$\hat{\alpha}^b_{1_0(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \kappa_{u,0,1,i}, \quad \hat{\alpha}^b_{1_0(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^n \xi_i^b \kappa_{u,0,0,i}.$$

From these bootstrap estimates of the average structural distribution functions, we obtain bootstrap LQTE estimates as above through inversion by linearly interpolating the value of the distribution function between the finite set of points at which we have estimated values to find the value of the outcome associated with each quantile in the set $q \in [0.1, 0.11, 0.12, \ldots, 0.89, 0.9]$, denoted $(\hat{\theta}_Y^{\leftarrow}(q,d))^b$. The bootstrap estimate of the LQTE for bootstrap replication b at point q is then $\hat{\Delta}^b(q) = (\hat{\theta}_Y^{\leftarrow}(q,1))^b - (\hat{\theta}_Y^{\leftarrow}(q,0))^b$. We form bootstrap standard error estimates for the LQTE at each quantile q as

$$s(q) = \big[q_{\mathrm{LQTE}}(0.75) - q_{\mathrm{LQTE}}(0.25)\big]/\big[q_N(0.75) - q_N(0.25)\big],$$

where $q_{\mathrm{LQTE}}(p)$ is the pth quantile of $\{\hat{\Delta}^b(q)\}_{b=1}^{500}$ and $q_N(p)$ is the pth quantile of the $N(0,1)$.

We also use the bootstrap LQTE estimates to obtain the critical values we use when plotting the uniform confidence bands in our example. We form bootstrap t-statistics for each quantile q as $t^b(q) = (\hat{\Delta}^b(q) - \hat{\Delta}(q))/s(q)$. We then take $t^b_{\max} = \max_q\{|t^b(q)|\}$ and use the 95th percentile of the bootstrap distribution of $t^b_{\max}$ as the critical value in constructing the confidence intervals for our figures following, for example, Chernozhukov, Fernández-Val, and Melly (2013).

REFERENCES

ABADIE, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292. [240]

——— (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [237,240,264-266]

AI, C., AND X. CHEN (2003): "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [238,256]

——— (2012): "The Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions," *Journal of Econometrics*, 170, 442–457. [238]

ANDREWS, D. W. K. (1994a): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. 4. Amsterdam: North-Holland, 2247–2294. [258]

——— (1994b): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72. [238]

ANGRIST, J. D., AND J.-S. PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. [234]

BACH, F. (2010): "Self-Concordant Analysis for Logistic Regression," *Electronic Journal of Statistics*, 4, 384–414. [239]

BELLONI, A., AND V. CHERNOZHUKOV (2011): "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130. [238,239,261]

——— (2013): "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19, 521–547. [239,261]

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. [235,237-239,244, 271,289]

BELLONI, A., V. CHERNOZHUKOV, I. FERNANDEZ-VAL, AND C. HANSEN (2017): "Supplement to 'Program Evaluation and Causal Inference With High-Dimensional Data'," *Econometrica Supplemental Material*, 85, http://dx.doi.org/10.3982/ECTA12723. [239]

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2010): "LASSO Methods for Gaussian Instrumental Variables Models," Preprint. Available at arXiv:1012.1297. [235,237,238]

——— (2013): "Inference for High-Dimensional Sparse Econometric Models," in *Advances in Economics and Econometrics. 10th World Congress of Econometric Society*, Vol. III. New York: Cambridge University Press, 245–295. [235,238,244,255]

——— (2014): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650. [235,238,249,253-255,261,266,271]

BELLONI, A., V. CHERNOZHUKOV, AND K. KATO (2013): "Uniform Post Selection Inference for LAD Regression Models," Preprint. Available at arXiv:1304.0282. [239,253]

BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98, 791–806. [261]

——— (2014): "Pivotal Estimation via Square-Root Lasso in Nonparametric Regression," *The Annals of Statistics*, 42, 757–788. [238]

BELLONI, A., V. CHERNOZHUKOV, AND Y. WEI (2013): "Honest Confidence Regions for Logistic Regression With a Large Number of Controls," Preprint. Available at arXiv:1304.3969. [239]

BENJAMIN, D. J. (2003): "Does 401(k) Eligibility Increase Saving? Evidence From Propensity Score Subclassification," *Journal of Public Economics*, 87, 1259–1290. [237,264-266]

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890. [236]

BICKEL, P. J. (1982): "On Adaptive Estimation," *The Annals of Statistics*, 10, 647–671. [237,238,258]

BICKEL, P. J., AND D. A. FREEDMAN (1981): "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217. [279]

BICKEL, P. J., Y. RITOV, AND A. B. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [239,261]

CANDÈS, E., AND T. TAO (2007): "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," *The Annals of Statistics*, 35, 2313–2351. [239]

CANER, M., AND H. H. ZHANG (2014): "Adaptive Elastic Net for Generalized Methods of Moments," *Journal of Business and Economic Statistics*, 32, 30–47. [239]

CATTANEO, M., M. JANSSON, AND W. NEWEY (2010): "Alternative Asymptotics and the Partially Linear Model With Many Regressors," Working Paper. Available at http://econ-www.mit.edu/files/6204. [268]

CATTANEO, M. D. (2010): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects Under Ignorability," *Journal of Econometrics*, 155, 138–154. [245]

CHAMBERLAIN, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596. [255,256]

CHAMBERLAIN, G., AND G. W. IMBENS (2003): "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18. [247]

CHEN, X. (2007): "Large Sample Sieve Estimatin of Semi-Nonparametric Models," in *Handbook of Econometrics*, Vol. 6. Amsterdam: North-Holland, 5559–5632. [234]

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of Semiparametric Models When the Criterion Function Is not Smooth," *Econometrica*, 71, 1591–1608. [238]

CHERNOZHUKOV, V., AND I. FERNÁNDEZ-VAL (2005): "Subsampling Inference on Quantile Regression Processes," *Sankhyā*, 67, 253–276. [248]

CHERNOZHUKOV, V., AND C. HANSEN (2004): "The Impact of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis," *Review of Economics and Statistics*, 86, 735–751. [237,264-266,271]

——— (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–262. [237]

——— (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132, 491–525. [237,247,258]

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, AND C. HANSEN (2016): "Double Machine Learning for Treatment and Causal Parameters". Preprint, arXiv:1608.00060. [237,258]

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2014): "Gaussian Approximation of Suprema of Empirical Processes," *The Annals of Statistics*, 42, 1564–1597. [281]

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268. [248,260,294]

CHERNOZHUKOV, V., C. HANSEN, AND M. SPINDLER (2015a): "Post-Selection and Post-Regularization Inference in Linear Models With Very Many Controls and Instruments," *American Economic Review: Papers and Proceedings*, 105, 486–490. [236,266]

——— (2015b): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [236]

CHESHER, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. [237]

DUDLEY, R. M. (1999): *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics, Vol. 63. Cambridge: Cambridge University Press. [272]

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. [265]

ENGEN, E. M., W. G. GALE, AND J. K. SCHOLZ (1996): "The Illusory Effects of Saving Incentives on Saving," *Journal of Economic Perspectives*, 10, 113–138. [265]

ESCANCIANO, J. C., AND L. ZHU (2013): "Set Inferences and Sensitivity Analysis in Semiparametric Conditionally Identified Models," CeMMAP Working Paper CWP55/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies. [237,258]

FAN, J., AND R. LI (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [239]

FARRELL, M. (2015): "Robust Inference on Average Treatment Effects With Possibly More Covariates Than Observations," *Journal of Econometrics*, 189, 1–23. [238]

FRANK, I. E., AND J. H. FRIEDMAN (1993): "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [239]

FRÖLICH, M., AND B. MELLY (2013): "Identification of Treatment Effects on the Treated With One-Sided Non-Compliance," *Econometric Reviews*, 32, 384–414. [241,266]

GINÉ, E., AND J. ZINN (1984): "Some Limit Theorems for Empirical Processes," *The Annals of Probability*, 12, 929–998. [246,247]

HAHN, J. (1997): "Bayesian Bootstrap of the Quantile Regression Estimator: A Large Sample Study," *International Economic Review*, 38, 795–808. [247]

——— (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [236,245]

HANSEN, B. E. (1996): "Inference When a Nuisance Parameter Is not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430. [247]

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [235]

HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286. [235]

HECKMAN, J. J., AND E. VYTLACIL (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96 (8), 4730–4734. [240]

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. [240]

HONG, H., AND D. NEKIPELOV (2010): "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1, 279–304. [241]

HONG, H., AND O. SCAILLET (2006): "A Fast Subsampling Method for Nonlinear Dynamic Models," *Journal of Econometrics*, 133, 557–578. [247]

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, 587–613. [239]
HUANG, J., J. L. HOROWITZ, AND F. WEI (2010): "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [239]
IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [240]
IMBENS, G. W., AND W. K. NEWEY (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. [237]
IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. [234]
KATO, K. (2011): "Group Lasso for High Dimensional Sparse Quantile Regression Models," Preprint. Available at arXiv:1103.1458. [239]
KLINE, P., AND A. SANTOS (2012): "A Score Based Approach to Wild Bootstrap Inference," *Journal of Econometric Methods*, 1, 23–41. [247]
KOENKER, R. (1988): "Asymptotic Theory and Econometric Practice," *Journal of Applied Econometrics*, 3, 139–147. [234]
——— (2005): *Quantile Regression*. Cambridge: Cambridge University Press. [242]
KOSOROK, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Berlin: Springer. [253,257]
LEEB, H., AND B. M. PÖTSCHER (2008a): "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338–376. [235,254]
——— (2008b): "Recent Developments in Model Selection and Related Areas," *Econometric Theory*, 24, 319–322. [235,254]
LINTON, O. (1996): "Edgeworth Approximation for MINPIN Estimators in Semiparametric Regression Models," *Econometric Theory*, 12, 30–60. [238]
MAMMEN, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21, 255–285. [247,266,267]
MEINSHAUSEN, N., AND B. YU (2009): "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," *The Annals of Statistics*, 37, 246–270. [239]
NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [238]
——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. [238]
——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168. [234]
NEYMAN, J. (1979): "$C(\alpha)$ Tests and Their Use," *Sankhyā*, 41, 1–21. [238]
OGBURN, E. L., A. ROTNITZKY, AND J. M. ROBINS (2015): "Doubly Robust Estimation of the Local Average Treatment Effect Curve," *Journal of the Royal Statistical Society, Series B*, 77, 373–396. [237,265]
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. [237,264,265]
——— (1995): "Do 401(k) Contributions Crowd out Other Personal Saving?" *Journal of Public Economics*, 58, 1–32. [237,264,265]
——— (1996): "Personal Retirement Saving Programs and Asset Accumulation: Reconciling the Evidence," Working Paper 5599, National Bureau of Economic Research. [237,264,265]
——— (2001): "The Transition to Personal Accounts and Increasing Retirement Wealth: Macro and Micro Evidence," Working Paper 8610, National Bureau of Economic Research. [237,264,265]
PÖTSCHER, B. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhyā*, 71, 1–18. [235,255]
ROBINS, J. M., AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [238]
ROBINSON, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [238]
ROMANO, J. P., AND A. M. SHAIKH (2012): "On the Uniform Asymptotic Validity of Subsampling and the Bootstrap," *The Annals of Statistics*, 40, 2798–2822. [237]
ROTHE, C., AND S. FIRPO (2013): "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions," Technical Report, NYU. [245]
SPINDLER, M., V. CHERNOZHUKOV, AND C. HANSEN (2016): "hdm: High-Dimensional Metrics," R Package Version 0.1.0. Available at http://CRAN.R-project.org/package=hdm. [233]
TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [239]
TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*. Berlin: Springer. [234]

VAN DE GEER, S. A. (2008): "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36, 614–645. [239]

VAN DER VAART, A. W. (1991): "On Differentiable Functionals," *The Annals of Statistics*, 19, 178–204. [238]

——— (1998): *Asymptotic Statistics*. Cambridge: Cambridge University Press. [253,257,258]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York: Springer. [246,247,249,253,257,272,274-276,279-281]

WASSERMAN, L. (2006): *All of Nonparametric Statistics*. New York: Springer. [234]

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data* (Second Ed.). Cambridge: MIT Press. [234]

ZOU, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [239]

Fuqua School of Business, Duke University, 100 Fuqua Drive, PO Box 90120, Office W312, Durham, NC 27708, U.S.A.; abn5@duke.edu,

Dept. of Economics, MIT, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, U.S.A.; vchern@mit.edu,

Dept. of Economics, Boston University, 270 Bay State Road, Room 415A, Boston, MA 02215-1403, U.S.A.; ivanf@bu.edu,

and

Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, U.S.A.; chansen1@chicagobooth.edu.

Co-editor Elie Tamer handled this manuscript.

Manuscript received 11 August, 2014; final version accepted 17 January, 2016; available online 24 March, 2016.