ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Chernozhukov, Victor; Semenova, Vira

Working Paper

Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions

cemmap working paper, No. CWP40/18

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Semenova, Vira (2018) : Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions, cemmap working paper, No. CWP40/18, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2018.4018

This Version is available at:
https://hdl.handle.net/10419/189761

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

cemmap
centre for microdata methods and practice

Simultaneous inference for Best Linear Predictor of the Conditional Average Treatment Effect and other structural functions

Victor Chernozhukov
Vira Semenova

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP40/18

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

Simultaneous Inference for Best Linear Predictor of the Conditional Average Treatment Effect and Other Structural Functions

Victor Chernozhukov, Vira Semenova
MIT
vchern@mit.edu, vsemen@mit.edu

Abstract

This paper provides estimation and inference methods for a structural function, such as Conditional Average Treatment Effect (CATE), based on modern machine learning (ML) tools. We assume that such function can be represented as a conditional expectation $g(x) = \mathbb{E}[Y_{\eta_0}|X = x]$ of a signal Y_{η_0}, where η_0 is the unknown nuisance function. In addition to CATE, examples of such functions include regression function with Partially Missing Outcome and Conditional Average Partial Derivative. We approximate $g(x)$ by a linear form $p(x)'\beta_0$, where $p(x)$ is a vector of the approximating functions and β_0 is the Best Linear Predictor. Plugging in the first-stage estimate $\widehat{\eta}$ into the signal $Y_{\widehat{\eta}}$, we estimate β_0 via ordinary least squares of $Y_{\widehat{\eta}}$ on $p(X)$. We deliver a high-quality estimate $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$, that features (a) a pointwise Gaussian approximation of $p(x_0)'\beta_0$ at a point x_0, (b) a simultaneous Gaussian approximation of $p(x)'\beta_0$ uniformly over x, and (c) optimal rate of convergence of $p(x)'\widehat{\beta}$ to $p(x)'\beta_0$ uniformly over x. In the case the misspecification error of the linear form decays sufficiently fast, these approximations automatically hold for the target function $g(x)$ instead of a pseudo-target $p(x)'\beta_0$. The first stage nuisance parameter η_0 is allowed to be high-dimensional and is estimated by modern ML tools, such as neural networks, l_1-shrinkage estimators, and random forest. Using our method, we estimate the average price elasticity conditional on income using Yatchew and No (2001) data and provide uniform confidence bands for the target regression function.

1 Introduction and Motivation

Many economic questions concern with a conditional average outcome

$$g(x) = \mathbb{E}[Y^*|X = x], \tag{1.1}$$

where Y^* is a latent variable of interest, $X \in \mathcal{X} \subset \mathcal{R}^r$ is a conditioning vector, and $g : \mathcal{X} \to \mathcal{R}$ is the target function. Examples of such functions include Conditional Average Treatment Effect (CATE), regression function with Partially Missing Outcome, and Conditional Average Partial Derivative (CAPD). Using additional identifying assumptions and/or additional data, a researcher constructs an observed signal Y_η indexed by a nuisance parameter η such that the true value of the signal $Y = Y_{\eta_0}$ is unbiased for $g(x)$:

$$g(x) = \mathbb{E}[Y|X = x].$$

In presence of multiple signals that are unbiased for $g(x)$ we focus on the signal Y_η that is robust. We call a signal Y_η robust if the pathwise derivative of its conditional on X expectation with respect to the nuisance parameter η is equal to zero:

$$\partial_\eta \mathbb{E}[Y_{\eta_0}|X = x][\eta - \eta_0] := \partial_r \mathbb{E}[Y_{\eta_0 + r(\eta - \eta_0)}|X = x] = 0 \quad \forall x \in \mathcal{X}.$$

If the signal Y_η is robust, its plug-in estimate $Y_{\widehat{\eta}}$ is insensitive to the biased estimation of $\widehat{\eta}$ and delivers a high-quality estimator of the target function $g(x)$ under mild conditions. In particular, the variance of $Y_{\widehat{\eta}}$ approximately equals to the variance of the infeasible signal Y_{η_0}, where the nuisance parameter η_0 is known. We say that a robust signal $Y_{\widehat{\eta}}$ delivers variance improvement over another unbiased signal Y^w_τ indexed by a different nuisance parameter τ if Y_η is robust and

$$Var(Y_{\widehat{\eta}}|X = x) \approx Var(Y_{\eta_0}|X = x) \leqslant Var(Y^w_{\tau_0}|X = x) \quad \forall x \in \mathcal{X}.$$

Consider the following examples in context of the regression function with Partially Missing Outcome. Define the Inverse Probability Weighting (IPW, Horwitz and Thompson (1952)) signal as

$$Y^w = \frac{DY^o}{s_0(Z)}$$

and a robust signal of Robins and Rotnitzky (1995) type:

$$Y := \mu_0(Z) + \frac{D}{s_0(Z)}[Y^o - \mu_0(Z)],$$

where $D \in \{1, 0\}$ is a binary indicator of the presence of Y^* in the sample, $Y^o = DY^*$ is the observed outcome, Z is a vector of observables such that the treatment status D is independent from the outcome Y^* and the covariate vector X conditionally on Z, $s_0(Z) = \mathbb{E}[D = 1|Z]$ is the

propensity score and $\mu_0(Z) = \mathbb{E}[Y^o|D=1,Z]$ is the conditional expectation function. The IPW signal Y_s^w is not robust to the biased estimation of the propensity score $s_0(Z)$:

$$\partial_s \mathbb{E}[Y_{s_0}^w|X] = -\mathbb{E}\frac{DY^*}{s_0^2(Z)}[s(Z)-s_0(Z)]|X \neq 0,$$

while the Robins and Rotnitzky type signal Y_η is robust to the biased estimation of $\eta_0(Z) = \{s_0(Z), \mu_0(Z)\}$:

$$\partial_{\eta_0}\mathbb{E}[Y_{\eta_0}|X] = \begin{bmatrix} -\mathbb{E}\frac{D}{s_0^2(Z)}[Y^o-\mu_0(Z)][s(Z)-s_0(Z)]|X \\ \mathbb{E}[1-\frac{D}{s_0(Z)}][\mu(Z)-\mu_0(Z)]|X \end{bmatrix} = 0.$$

Consequently, the bias in the estimation error of the propensity score $\widehat{s}(Z) - s_0(Z)$ translates into the bias of the estimated signal $Y_{\widehat{s}}^w$, but does not translate into such bias of the estimated robust signal $Y_{\widehat{\eta}}$. As a result, the estimate of the target function $g(x)$ based on $Y_{\widehat{\eta}}$ is high-quality and the one based on $Y_{\widehat{s}}^w$ is low-quality. In addition to the robustness comparison of Y_η and Y_s^w, the signal $Y_{\widehat{\eta}}$ delivers variance improvement over $Y_{s_0}^w$:

$$Var(Y_{\widehat{\eta}}|X) \approx Var(Y_{\eta_0}|X) \leqslant Var(Y_{s_0}^w|X).$$

Therefore, a robust signal Y_η is preferred to a non-robust signal Y_s^w when $s_0(Z)$ is unknown due to robustness and when $s_0(Z)$ is known due to variance reduction.

Assuming a robust signal Y is available, we approximate the target function $g(x)$ at a point x by a linear form $p(x)'\beta_0$:

$$g(x) = p(x)'\beta_0 + r_g(x),$$

where $p(x)$ is a d-vector of technical transformations of the covariates x, $r_g(x)$ is the misspecification error due to the linear approximation[1], and β_0 is the Best Linear Predictor, defined by the balancing equation:

$$\mathbb{E}p(X)[g(X) - p(X)'\beta_0] = \mathbb{E}p(X)r_g(X) = 0.$$

The two-stage estimator $\widehat{\beta}$ of Best Linear Predictor β_0, which we refer to as Locally Robust estimator, is constructed as follows. In the first stage we construct an estimate $\widehat{\eta}$ of the nuisance parameter η_0. In the second stage we construct an estimate $\widehat{Y}_i$ of the signal Y_i as $\widehat{Y}_i := Y_i(\widehat{\eta})$ and run ordinary least squares of $\widehat{Y}_i$ on the technical regressors $p(X_i)$. We use different samples for the estimation of η in the first stage and the estimation of β_0 in the second stage in a form of cross-fitting, described in the following definition.

Definition 1.1 (Cross-fitting). *1. For a random sample of size N, denote a K-fold random partition of the sample indices $[N] = \{1, 2, ..., N\}$ by $(J_k)_{k=1}^K$, where K is the number of partitions and sample size of each fold is $n = N/K$. Also for each $k \in [K] = \{1, 2, ..., K\}$*

[1] Our analysis allows for either vanishing and non-vanishing specification error $r_g(x)$.

define $J_k^c = \{1, 2, ..., N\} \setminus J_k$.

2. *For each* $k \in [K]$, *construct an estimator* $\widehat{\eta}_k = \widehat{\eta}(V_{i \in J_k^c})$[2] *of the nuisance parameter value* η_0 *using only the data from* J_k^c. *For any observation* $i \in J_k$, *define an estimated signal* $\widehat{Y}_i := Y_i(\widehat{\eta}_k)$.

Definition 1.2 (Locally Robust Estimator)**.** *Given an estimate of a signal* $(\widehat{Y}_i)_{i=1}^N$, *define Locally Robust Estimator as:*

$$\widehat{\beta} := \{\frac{1}{N}\sum_{i=1}^{N} p(X_i)p(X_i)'\}^{-1}\frac{1}{N}\sum_{i=1}^{N} p(X_i)\widehat{Y}_i. \tag{1.2}$$

Under the mild conditions on η, Locally Robust delivers a high-quality estimate $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$ with the following properties:

- W.p. $\to 1$, the mean squared error of $p(x)'\widehat{\beta}$ is bounded by

$$(\mathbb{E}_N(p(X_i)'(\widehat{\beta} - \beta_0))^2)^{1/2} = O_P(\sqrt{\frac{d}{N}}) + z_d,$$

  where z_d is the effect of the misspecification error $r_g(x)$.

- The estimator $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$ is asymptotically linear:

$$\sqrt{N}\frac{p(x)'(\widehat{\beta} - \beta_0)}{\sqrt{p(x)'\Omega p(x)}} = G_N(x) + o_P(1),$$

  where the empirical process $G_N(x)$ converges to a tight Gaussian process with marginal distribution $N(0, 1)$ uniformly over $x \in \mathcal{X}$ and the covariance matrix Ω can be consistently estimated by a sample analog $\widehat{\Omega}$.

- In the case the misspecification error $r_g(x)$ is small, the pseudo-target function $p(x)'\beta_0$ can be replaced by the target function $g(x)$:

$$\sqrt{N}\frac{p(x)'\widehat{\beta} - g(x)}{\sqrt{p(x)'\Omega p(x)}} = G_N(x) + o_P(1).$$

The results of this paper accommodate the estimation of $\widehat{\eta}$ by high-dimensional/highly complex modern machine learning (ML) methods, such as random forests, neural networks, and l_1-shrinkage estimators, as well as earlier developed tools. The only requirement we impose on the estimation of $\widehat{\eta}$ is its mean square convergence to the true nuisance parameter η_0 at a high-quality rate $o_P(N^{-1/4-\delta}/\sup_{x\in\mathcal{X}} \|p(x)\|^{1/2}), \delta > 0$. Under suitably chosen growth rate of the dimension of the basis functions $d = d(N)$, this requirement is satisfied under structured assumptions on

[2]The results of this paper hold without relying on the specific choice of the first stage estimator $\widehat{\eta}(Z)$. For the possible ways to estimate η_0, see the discussion.

η_0, such as approximate sparsity of η_0 with respect to some dictionary, well-approximability of η_0 by trees or by sparse neural and deep neural nets.

1.1 Examples

Examples below apply the proposed framework to study Conditional Average Treatment Effect, regression function with Partially Missing Outcome and Conditional Average Partial Derivative.

Example 1 (Conditional Average Treatment Effect)**.** Let Y^1 and Y^0 be the potential outcomes, corresponding to the response of a subject with and without receiving a treatment, respectively. Let $D \in \{1, 0\}$ indicate the subject's presence in the treatment group. The object of interest is the Conditional Average Treatment Effect

$$g(x) := \mathbb{E}[Y^1 - Y^0 | X = x].$$

Since an individual cannot be treated and non-treated at the same time, only the actual outcome $Y^o = DY^1 + (1 - D)Y^0$, but not the treatment effect $Y^1 - Y^0$, is observed.

A standard way to make progress in this problem is to assume unconfoundedness (Rosenbaum and Rubin (1983)). Suppose there exists an observable control vector Z such that the treatment status D is independent of the potential outcomes Y^1, Y^0 and the covariates X conditionally on Z.

Assumption 1.1 (Unconfoundedness)**.** *The treatment status D is independent of the potential outcomes Y^1, Y^0 conditionally on Z: $\{Y^1, Y^0\} \perp D|Z$.*

Assumption 1.2 (Random Treatment Assignment)**.** *The treatment status D is independent of X conditionally on Z: $\mathbb{E}[D = 1|X, Z] = \mathbb{E}[D = 1|Z]$.*

Define the conditional probability of treatment receipt as $s_0(Z) = \mathbb{E}[D = 1|Z]$. An immediate feasible choice of an unbiased signal Y^w is the difference between the realized outcomes in the treatment DY^o and the control groups $(1 - D)Y^o$, inversely weighted by the probability of presence of in the respective group:

$$Y^w := \frac{DY^o}{s_0(Z)} - \frac{(1 - D)Y^o}{1 - s_0(Z)} = \frac{D - s_0(Z)}{s_0(Z)(1 - s_0(Z)} Y^o.$$

To show that the signal is unbiased:

$$\mathbb{E}[Y^w|X] = g(x),$$

recognize that a stronger statement $\mathbb{E}[Y^w|Z,X] = \mathbb{E}[Y^1 - Y^0|X,Z]$ holds:

$$\begin{aligned}
\mathbb{E}[Y^w|Z,X] &= \frac{\mathbb{E}[D=1|Z,X]\mathbb{E}[Y^1|D=1,X,Z]}{s_0(Z)} - \frac{\mathbb{E}[D=0|Z,X]\mathbb{E}[Y^0|D=0,X,Z]}{1-s_0(Z)} \\
&= \mathbb{E}[Y^1|D=1,X,Z] - \mathbb{E}[Y^0|D=0,X,Z] && \text{(Assumption 1.2)} \\
&= \mathbb{E}[Y^1 - Y^0|X,Z]. && \text{(Assumption 1.1)}
\end{aligned}$$

Although the feasible signal Y^w is unbiased, it can be improved in terms of the robustness with respect to biased estimation of $s_0(Z)$ and the noise reduction. Consider a robust signal Y of Robins and Rotnitzky (1995) type[3] :

$$Y := \mu_0(1,Z) - \mu_0(0,Z) + \frac{D[Y^o - \mu_0(1,Z)]}{s_0(Z)} - \frac{(1-D)[Y^0 - \mu_0(0,Z)]}{1-s_0(Z)}, \tag{1.3}$$

where $\mu_0(D,Z) = \mathbb{E}[Y^o|D,Z]$ is the conditional expectation function of Y^o given D, Z. Corollary 3.1 shows that the signal Y is robust to the estimation error of the nuisance parameter $\eta(Z) := (s(Z), \mu(1,Z), \mu(0,Z))$. Lemma 6.5 shows that the signal Y achieves lower conditional variance than the signal Y^w:

$$Var(Y_{\widehat{\eta}}|X=x) \approx Var(Y_{\eta_0}|X=x) \leqslant Var(Y^w_{s_0}|X=x) \quad \forall x \in \mathcal{X}.$$

Remark 1.1 (Experimental Data)**.** In case a researcher assigns the treatment status D at random conditional on a vector of stratification variables Z_D, Assumptions 1.1 and 1.2 hold by construction of D for $Z = Z_D$. Although the relationship between X and Z is no longer restricted, one may include X and other observables into the control vector Z to reduce noise of the estimate of $g(x)$.

Example 2 (Regression Function with Partially Missing Outcome)**.** Suppose a researcher is interested in the conditional expectation given a covariate vector X of a variable Y^*

$$g(x) := \mathbb{E}[Y^*|X=x]$$

that is partially missing. Let $D \in \{1,0\}$ indicate whether the outcome Y^* is observed, and $Y^o = DY^*$ be the observed outcome. Since the researcher does not control the presence status D, a standard way to make progress is to assume existence of an observable control vector Z such that $Y \perp D|Z$.

Assumption 1.3 (Missingnesss at Random)**.** *The presence indicator D is independent from the outcome Y conditionally on Z: $Y \perp D|Z$.*

Since setting Z to a full vector of observables (in particular, having X as part of Z) makes the assumption 1.3 the least restrictive, we will assume $X \subseteq Z$ within the context of the example.

[3]This signal is also doubly robust in the sense of Robins et al. (1994), which is a different notion of robustness, motivated by misspecification of the nuisance parameter η.

Let the conditional probability of presence be

$$s_0(Z) = \mathbb{E}[D = 1|Z, Y^o] = \mathbb{E}[D = 1|Z].$$

An immediate feasible choice of the unbiased signal is the observed outcome $Y^o = DY^*$ inversely weighted by $s_0(Z)$:

$$Y^w := \frac{DY^*}{s_0(Z)}.$$

That Y^w is an unbiased signal for $g(x)$:

$$\mathbb{E}[Y^w|X = x] = \mathbb{E}[Y^*|X = x],$$

follows from a stronger statement:

$$\mathbb{E}[Y^w|Z = z] = \frac{s_0(Z)\mathbb{E}[Y^o|D = 1, Z = z]}{s_0(Z)} = \mathbb{E}[Y^*|Z = z],$$

where the last equality follows from the Missingnesss at Random. Since $X \subseteq Z$, the desired statement follows.

Although the feasible signal Y^w is unbiased, it can be improved in terms of the robustness with respect to biased estimation of $s_0(Z)$ and the noise reduction. Consider the robust signal Y:

$$Y = \mu_0(Z) + \frac{D[Y^o - \mu_0(Z)]}{s_0(Z)}, \tag{1.4}$$

where the function $\mu_0(Z) = \mathbb{E}[Y^o|Z] = \mathbb{E}[Y|Z, D = 1]$ is the conditional expectation function of the observed outcome Y^o given Z and $D = 1$. Corollary 3.2 shows that the signal Y is robust to the estimation error of the nuisance parameter $\eta(Z) := (s(Z), \mu(1, Z), \mu(0, Z))$. Lemma 6.4 shows that the signal Y achieves lower conditional variance than the signal Y^w:

$$Var(Y_{\widehat{\eta}}|X = x) \approx Var(Y_{s_0}|X = x) \leqslant Var(Y^w_{\eta_0}|X = x) \quad \forall x \in \mathcal{X}.$$

Example 3 (Experiment with Partially Missing Outcome)**.** Let Y^1 and Y^0 be the potential outcomes, corresponding to the response of a subject with and without receiving a treatment, respectively. Suppose a researcher is interested in the Conditional Average Treatment Effect

$$g(x) := \mathbb{E}[Y^1 - Y^0|X = x].$$

Conditionally on a vector of stratifying variables Z_D, he randomly assigns the treatment status $T \in \{1, 0\}$ to measure the outcome $TY^1 + (1-T)Y^0$. In presence of the partially missing outcome, let $D \in \{1, 0\}$ indicate the presence of the outcome record $Y^o = D(TY^1 + (1 - T)Y^0)$ in the data. Since the presence indicator D may be co-determined with the covariates X, estimating the treatment effect function $g(x)$ on the observed outcomes only without accounting for the

Missingnesss may lead to an inconsistent estimate of $g(x)$.

Since the researcher does not control presence status D, a standard way to make progress is to assume Missingnesss at Random, namely existence of an observable control vector Z such that $Y \perp D|Z$. Setting Z to be a full vector of observables (in particular, $X \subseteq Z$) makes Missingness at Random (Assumption 1.3) the least restrictive. Define the conditional probability of presence

$$s_0(Z,T) = \mathbb{E}[D = 1|Y^o, Z, T] = \mathbb{E}[D = 1|Z, T]$$

and the treatment propensity score

$$h_0(Z) := \mathbb{E}[T = 1|Z].$$

A robust signal Y for the CATE $g(x)$ can be obtained as follows:

$$Y = \mu_0(1,Z) - \mu_0(0,Z) + \frac{DT[Y^o - \mu_0(1,Z)]}{s_0(Z,T)h_0(Z)} - \frac{D(1-T)[Y^o - \mu_0(0,Z)]}{s_0(Z,T)(1-h_0(Z))}, \tag{1.5}$$

where $\mu_0(T,Z) = \mathbb{E}[Y^o|T,Z]$ is the conditional expectation function of Y^o given T, Z. That Y is an unbiased signal for $g(x)$:

$$\mathbb{E}[Y|Z] = \mathbb{E}[Y^1 - Y^0|Z]$$

follows from a stronger statement:

$$\begin{aligned}\mathbb{E}[Y|Z] &= \mu_0(1,Z) - \mu_0(0,Z) + \frac{s_0(Z,T)\mathbb{E}[TY^o|T=1,Z]}{s_0(Z,T)h_0(Z)} - \frac{s_0(Z,T)\mathbb{E}[(1-T)Y^o|T=0,Z]}{s_0(Z,T)(1-h_0(Z))} \\ &\quad - \mu_0(1,Z) + \mu_0(0,Z) \\ &= \mathbb{E}[Y^1 - Y^0|Z],\end{aligned}$$

where the last equality follows from Assumption 1.3. Using the arguments in the Corollaries 3.1 and 3.2, it can be shown that the nuisance parameter $\eta = \{\mu(T,Z), s(Z,T)\}$ consists of the conditional expectation function $\mu_0(T,Z)$ and the propensity score $s_0(Z,T)$.

Example 4 (Conditional Average Partial Derivative)**.** Let

$$\mu(x,w) := \mathbb{E}[Y^o|X = x, W = w]$$

be a conditional expectation function of an outcome Y^o given a set of variables X, W. Suppose a researcher is interested in the conditional average derivative of $\mu(x,w)$ with respect to w given $X = x$, denoted by

$$Y^* := \partial_w \mu(X,w)|_{w=W}.$$

An immediate choice of the unbiased signal Y^w for the latent variable Y^* follows from integration

by parts. Specifically,

$$\begin{aligned} g(x) &= \mathbb{E}[\partial_w \mu(x, W)|X = x] \\ &= -\mathbb{E}[\mu(x, w)\partial_w \log f(W = w|X = x)|X = x]. \end{aligned}$$

Therefore,

$$Y^w := -Y^o \partial_w \log f(W|X)$$

is an unbiased signal for $\partial_w \mu(x, W)$. We consider a robust signal Y of Newey and Stoker (1993) type:

$$Y := -\partial_w \log f(W|X)[Y^o - \mu(X, W)] + \partial_w \mu(X, W),$$

where $f(W|X = x)$ is the conditional density of W conditionally on $X = x$. To see that Y is an unbiased signal for $g(x)$, recognize that

$$\mathbb{E}[Y|X, W] = -\partial_w \log f(W|X)[\mathbb{E}[Y^o|X, W] - \mu(X, W)] + \partial_w \mu(X, W) = \partial_w \mu(X, W) \tag{1.6}$$

by definition of $\mu(x, w)$. The nuisance parameter $\eta = \{\mu(X, W), f(W|X)\}$ consists of the conditional expectation function $\mu(X, W)$ and the conditional density $f(W|X)$. Corollary 3.3 shows that the signal Y is robust to the estimation error of the nuisance parameter $\eta(X, W) = \{\mu(X, W), f(W|X)\}$.

1.2 Literature Review

This paper builds on the three bodies of research within the semiparametric literature: orthogonal(debiased) machine learning, least squares series estimation, and treatment effects/missing data problems. The first literature provides a $\sqrt{N}$-consistent and asymptotically normal estimates of low-dimensional target parameters in the presence of high-dimensional/highly complex nonparametric nuisance functions. The second one provides the pointwise and uniform limit theory for least squares series estimator. The third one provides the efficiency bounds in various problems concerned with missing data or treatment effects.

Orthogonal machine learning (Chernozhukov et al. (2016a), Chernozhukov et al. (2016b)) concerns with the inference on a fixed-dimensional target parameter β in presence of a high-dimensional nuisance function η in a semiparametric moment problem. In case the moment condition is Neyman orthogonal (robust) to the perturbations of η, the estimation of η by ML methods has no first-order effect on the asymptotic distribution of the target parameter β. In particular, plugging in an estimate of η obtained on a separate sample, results in a $\sqrt{N}$-consistent asymptotically normal estimate whose asymptotic variance is the same as if $\eta = \eta_0$ was known. This result allows one to use highly complex machine learning methods to estimate the nuisance function η, such as l_1 penalized methods in sparse models (Bühlmann and van der Geer (2011), Belloni et al. (2016)), L_2 boosting in sparse linear models (Luo and Spindler (2016)), and other

methods for classes of neural nets, regression trees, and random forests. We extend this result in two directions: (1) we allow the dimension of β to grow with sample size and (2) we provide a simultaneous approximation of the pseudo-target function $p(x)'\beta$ by a Gaussian process.

The second building block of our paper is the literature on least squares series estimation (Newey (2007), Belloni et al. (2015)), which establishes the pointwise and the uniform limit theory for least squares series estimation. We extend this theory by allowing the outcome variable Y to depend upon an unknown nuisance parameter η, and do so without adding any assumptions on the problem design.

The third relevant body of literature are the efficiency bounds in missing data and treatment effects problems. In the class of general missing data models, efficiency bounds have been established under the assumption that the propensity score can be modeled by a finite-dimensional parameter(Graham (2011), Graham et al. (2012)), which is a restrictive condition. As for the efficiency bounds for the average treatment effect (Hahn (1998), Hirano et al. (2003)), these papers rely on the estimation of the propensity score (or other nuisance parameters) by kernel or series methods, which fail in modern high-dimensional settings. By combining robustness with sample splitting, we relax the P-Donsker requirement on the propensity score and allow it to be estimated by high-dimensional/highly complex machine learning methods. Other examples of using machine learning for the estimation of the treatment effects include Wager and Athey (2016), which provides a pointwise Gaussian approximation to a Conditional Average Treatment Effect using random forest in the classical low-dimensional setting.

2 Asymptotic Theory

We shall use empirical process notation. For a generic function f and a generic sample $(X_i)_{i=1}^N$, denote a sample average by

$$\mathbb{E}_N f(X_i) := \frac{1}{N}\sum_{i=1}^N f(X_i)$$

and a $\sqrt{N}$-scaled, demeaned sample average by

$$\mathbb{G}_N f(X_i) := \frac{1}{\sqrt{N}}\sum_{i=1}^N (f(X_i) - \mathbb{E}f(X_i)).$$

All asymptotic statements below are with respect to $N \to \infty$.

Assumption 2.1 (Identification)**.** *Let $Q := \mathbb{E}p(X)p(X)'$ denote population covariance matrix of technical regressors. Assume that $\exists \quad 0 < C_{\min} < C_{\max} < \infty$ s.t. $C_{\min} < \min \operatorname{eig}(Q) < \max \operatorname{eig}(Q) < C_{\max}$.*

Assumption 2.1 requires that the regressors $p(X)$ are not too collinear in population, which allows identification of the best linear predictor β_0.

Assumption 2.2 (Growth Condition)**.** *We assume that the* sup*-norm of the technical regressors* $\xi_d^2 := \sup_{x\in\mathcal{X}} \|p(x)\| = \sup_{x\in\mathcal{X}}(\sum_{j=1}^d p_j(x)^2)^{1/2}$ *grows sufficiently slow:*

$$\sqrt{\frac{\xi_d^2 \log N}{N}} = o(1).$$

Assumption 2.3 (Misspecification Error)**.** *There exists a sequence of finite constants* $l_d, r_{d\,d\to\infty}$ *such that the norms of the misspecification error are controlled as follows:*

$$\|r_g\|_{F,2} := \sqrt{\int r_g(x)^2(x)dF(x)} \lesssim r_d \text{ and } \|r_g\|_{F,\infty} := \sup_{x\in\mathcal{X}} |r_g(x)| \lesssim l_d r_d.$$

Assumption 2.3 introduces the rate of decay of the misspecification error. Specifically, the sequence of constants r_d bounds the mean squared misspecification error. In addition, the sequence $l_d r_d$ bounds the worst-case misspecification error uniformly over the domain of X $\mathcal{X}$, where l_d is the modulus of continuity of the worst-case error with respect to mean squared error.

Define the sampling error U as follows:

$$U := Y - g(X).$$

Assumption 2.4 (Sampling Error)**.** *The second moment of the sampling error* U *conditionally on* X *is bounded from above by* $\bar{\sigma}$*:*

$$\sup_{x\in\mathcal{X}} \mathbb{E}[U^2|X=x] \lesssim_P \bar{\sigma}^2.$$

Assumption 2.5 (Small Bias Condition)**.** *There exists a sequence* $\epsilon_N = o(1)$*, such that with probability at least* $1-\epsilon_N$*, the first stage estimate* $\widehat{\eta}$*, obtained by cross-fitting (Definition 1.1), belongs to a shrinking neighborhood of* η_0*, denoted by* T_N*. Uniformly over* T_N*, the following mean square convergence holds:*

$$B_N := \sqrt{N} \sup_{\eta\in T_N} \|\mathbb{E}p(X)[Y_\eta - Y_{\eta_0}]\| = o(1), \tag{2.1}$$

$$\xi_d \kappa_N := \xi_d \sup_{\eta\in T_N} (\mathbb{E}(Y_\eta - Y_{\eta_0})^2)^{1/2} = o(1). \tag{2.2}$$

Assumption 2.5 introduces the realization set T_N, where the first stage estimate $\widehat{\eta}$ belongs to with probability approaching one. It provides a restriction jointly on the speed of shrinkage of the set T_N (and, hence, the quality of the first stage estimate $\widehat{\eta}$), robustness of the unobserved signal Y_η with respect to the biased estimation of η, and the growth speed ξ_d of the technical regressors. Section 3 shows that the Assumption 2.5 holds for Examples 1, 2, 4.

Assumption 2.6 (Tail Bounds)**.** *There exist* $m > 2$ *such that the upper bound of the* m*'th*

moment of $|U|$ is bounded conditionally on X:

$$\sup_{x\in\mathcal{X}} \mathbb{E}[|U|^m|X=x] \lesssim 1.$$

The norm of the outer product of the technical regressors grows sufficiently slow:

$$\max_{1\leqslant i\leqslant N} \|p_i p_i'\| \frac{d}{N} = o(1).$$

Assumption 2.6 bounds the tail of the distribution of the sampling error U and the regressors $p(X)$.

Assumption 2.7 (Bound on Regression Errors)**.** *There exists a sequence $\epsilon_N = o(1)$ and a constant $q > 2$, such that with probability at least $1-\epsilon_N$, the first stage estimate $\widehat{\eta}$, obtained by cross-fitting (Definition 1.1), belongs to a shrinking neighborhood of η_0, denoted by T_N, constrained as follows:*

$$\sup_{\eta\in T_N} (\mathbb{E}|Y_\eta - Y_{\eta_0}|^q)^{1/q} = O(1).$$

2.1 Pointwise Limit Theory

Theorem 2.1 (Pointwise Limit Theory of LRE)**.** *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5 hold. Then, the following statements hold:*

(a) The second norm of the estimation error is bounded as:

$$\|\widehat{\beta}-\beta_0\|_2 \lesssim_P \sqrt{\frac{d}{N}} + \big[\sqrt{d} l_d c_d \wedge \xi_d c_d/\sqrt{N}\big],$$

which implies a bound on the mean squared error of the estimate $p(x)'\widehat{\beta}$ of the pseudo-target function $p(x)'\beta_0$:

$$(\mathbb{E}_N(p(X_i)'(\widehat{\beta}-\beta_0))^2)^{1/2} \lesssim_P \sqrt{\frac{d}{N}} + \big[\sqrt{d} l_d c_d \wedge \xi_d c_d/\sqrt{N}\big].$$

(b) For any $\alpha \in \mathcal{S}^{d-1} := \{\alpha \in \mathcal{R}^d : \|\alpha\| = 1\}$ the estimator $\widehat{\beta}$ is approximately linear:

$$\sqrt{N}\alpha'(\widehat{\beta}-\beta) = \alpha' Q^{-1}\mathbb{G}_N p(X_i)(U_i + r_g(X_i)) + R_{1,N}(\alpha),$$

where the remainder term $R_{1,N}(\alpha)$ is bounded as $R_{1,N}(\alpha) \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}(1+\sqrt{d} l_d c_d) + \xi_d \kappa_N$.

(c) Define the asymptotic covariance matrix of the $p(x)'(\widehat{\beta}-\beta_0)$ as follows:

$$\Omega = Q^{-1}\mathbb{E}p(X)p(X)'(U+r_g(X))^2 Q^{-1}.$$

If $R_{1,N}(\alpha) = o_P(1)$ and the Lindeberg condition holds: $\lim_{M\to\infty} \mathbb{E}U^2 1_{|U|>M} \to 0$, then the pointwise estimator is approximately Gaussian:

$$\lim_{N\to\infty} \sup_{t\in\mathcal{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}\alpha'(\widehat{\beta} - \beta_0)}{\sqrt{\alpha'\Omega\alpha}} < t \right) - \Phi(t) \right| = 0. \tag{2.3}$$

In particular, for any point $x_0 \in \mathcal{X}$ for $\alpha = \frac{p(x_0)}{\|p(x_0)\|}$, the estimator $p(x_0)'\widehat{\beta}$ of the pseudo-target value $p(x_0)'\beta_0$ is asymptotically normal:

$$\lim_{N\to\infty} \sup_{t\in\mathcal{R}} \left| \mathrm{P}\left(\frac{\sqrt{N}p(x_0)'(\widehat{\beta} - \beta_0)}{\sqrt{p(x_0)'\Omega p(x_0)}} < t \right) - \Phi(t) \right| = 0. \tag{2.4}$$

(d) In addition, if Assumptions 2.6 and 2.7 hold, Ω can be consistently estimated by a sample analog:

$$\widehat{\Omega} := \widehat{Q}^{-1}\mathbb{E}_N p(X_i)p(X_i)'(Y_i(\widehat{\eta}) - p(X_i)'\widehat{\beta})^2\widehat{Q}^{-1}. \tag{2.5}$$

Theorem 2.1 is our first main result. Under small bias condition, Locally Robust Estimator has the oracle rate, oracle asymptotic linearity representation and asymptotic variance Ω, where the oracle knows the true value of the first-stage nuisance parameter η_0.

2.2 Uniform Limit Theory

Let $\alpha(x) := p(x)/\|p(x)\|$ denote the normalized value of technical regressors $p(x)$. Define their Lipshitz constant as:

$$\xi_d^L = \sup_{x,x'\in\mathcal{X}, x\neq x'} \frac{\|\alpha(x) - \alpha(x')\|}{\|x - x'\|}.$$

Assumption 2.8 (Basis)**.** *Basis functions are well-behaved, namely (i) $\frac{(\xi_d^L)^{2m/(m-2)}\log N}{N} \lesssim 1$ and $\log \xi_d^L \lesssim \log d$.*

Assumption 2.9 (First Stage Error Bound and Rate)**.** *There exists a sequence $\epsilon_N = o(1)$, such that with probability at least $1-\epsilon_N$, the first stage estimate $\widehat{\eta}$, obtained by cross-fitting (Definition 1.1), belongs to a shrinking neighborhood of η_0, denoted by T_N, constrained as follows:*

$$\xi_d \sup_{\eta\in T_N} (\mathbb{E}(Y_\eta - Y_{\eta_0})^2)^{1/2}\sqrt{\log N} = o(1).$$

Theorem 2.2 (Uniform Limit Theory of LRE)**.** *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.8, 2.9 hold.*

(a) The estimator is approximately linear uniformly over the domain $\mathcal{X}$:

$$|\sqrt{N}\alpha(x)'(\widehat{\beta} - \beta_0) - \alpha'(x)\mathbb{G}_N p(X_i)[U_i + r_g(X_i)]| \leqslant R_{1,N}(\alpha(x))$$

where $R_{1,N}(\alpha(x))$, summarizing the impact of unknown design and the first stage misspecification error, obeys

$$\sup_{x\in\mathcal{X}} R_{1,N}(\alpha(x)) \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}(N^{1/m}\sqrt{\log N} + \sqrt{d} l_d r_d) + \\ \xi_d \kappa_N \sqrt{\log N} + N^{-1/2+1/q}\log N =: \bar{R}_{1N}$$

uniformly over $x \in \mathcal{X}$. Moreover,

$$|\sqrt{N}\alpha(x)'(\widehat{\beta} - \beta_0) - \alpha'(x)\mathbb{G}_N p(X_i)[U_i + r_g(X_i)]| \leqslant R_{1,N}(\alpha(x)) + R_{2,N}(\alpha(x))$$

where $R_{2,N}(\alpha(x))$, summarizing the impact of misspecification error, obeys

$$R_{2,N}(\alpha(x)) \lesssim_P \sqrt{\log N} l_d r_d =: \bar{R}_{2N}$$

uniformly over $x \in \mathcal{X}$.

(b) The estimator $p(x)'\widehat{\beta}$ of the pseudo-target $p(x)'\beta_0$ converges uniformly over $\mathcal{X}$:

$$\sup_{x\in\mathcal{X}} |p(x)'(\widehat{\beta} - \beta_0)| \lesssim_P \frac{\xi_d}{\sqrt{N}}[\sqrt{\log N} + \bar{R}_{1N} + \bar{R}_{2N}].$$

Theorem 2.2 is our second main result in the paper. Under small bias condition, Locally Robust Estimator achieves oracle asymptotic linearity representation uniformly over the domain $\mathcal{X} \subset \mathcal{R}^r$ of the covariates of interest X.

Remark 2.1 (Optimal Uniform Rate in Holder class)**.** Suppose the true function g belongs to the Holder smoothness class of order k, denoted by $\Sigma_k(\mathcal{X})$. Then, the optimal number d of technical regressors that comprise a vector $p(x)$ obeys

$$d \asymp (\log N/N)^{-r/(2k+r)}.$$

This choice of d yields the optimal uniform rate:

$$\sup_{x\in\mathcal{X}} |\widehat{g}(x) - g(x)| \lesssim_P \left(\frac{\log N}{N}\right)^{r/(2k+r)}.$$

Our result on strong approximation by a Gaussian process plays an important role in our second result on inference that is concerned with the weighted bootstrap. Consider a set of weights $h_1, h_2, \ldots, h_N$ that are i.i.d. draws from the standard exponential distribution and are independent of the data. For each draw of such weights, define the weighted bootstrap draw of the least squares estimator as a solution to the least squares problem weighted by $h_1, h_2, \ldots, h_N$, namely

$$\widehat{\beta}^b \in \arg\min_{b\in\mathcal{R}^k} \mathbb{E}_N[h_i(\widehat{Y}_i - p(X_i)'b)^2].$$

For all $x \in \mathcal{X}$, define $\widehat{g}^b(x) = p(x)\widehat{\beta}^b$. The following corollary establishes validity of weighted bootstrap for approximating the distribution of series process.

Corollary 2.1 (Weighted Bootstrap Method). *(a) Let Assumption 2.6 be satisfied with $m \geqslant 3$. In addition, assume that $\bar{R}_{1,N} = o_P(a_N^{-1})$ and $a_N^6 d^4 \xi_d^2 (1 + l_d^3 c_d^3)^2 \log^2 N/N \to 0$. Then, for some $\mathcal{N}_d \sim N(0, J_k)$*

$$\sqrt{N}\frac{\alpha(x)'(\widehat{\beta} - \beta_0)}{\|\alpha(x)\Omega^{1/2}\|} =_d \frac{\alpha(x)'\Omega^{1/2}}{\|\alpha(x)\Omega^{1/2}\|}\mathcal{N}_d + o_P(a_N^{-1})$$

in $l^\infty(\mathcal{X})$ so that for $e(x) = \Omega^{1/2}p(x)$,

$$\sqrt{N}\frac{p(x)'(\widehat{\beta} - \beta_0)}{\|e(x)\|} =_d \frac{e(x)}{\|e(x)\|}\mathcal{N}_d + o_P(a_N^{-1})$$

in $l^\infty(\mathcal{X})$.

(b) The weighted bootstrap process satisfies:

$$\sqrt{N}\alpha(x)'(\widehat{\beta}^b - \widehat{\beta}) = \alpha(x)'\mathbb{G}_N[(h_i - 1)p(X_i)(U_i + r_g(X_i)] + R_{1N}^b(\alpha(x)),$$

where the remainder obeys

$$R_{1N}^b(\alpha(x)) \lesssim_P \sqrt{\frac{\xi_d^2 \log^3 N}{N}}(N^{1/m}\sqrt{\log N} + \sqrt{d} l_d r_d) = o(1/\log N) =: \bar{R}_{1N}^b$$

(c) $\sqrt{N}\frac{p(x)'(\widehat{\beta}^b - \widehat{\beta})}{\|e(x)\|} =^d \frac{e(x)'}{\|e(x)\|}\mathcal{N}_d + o_P(a_N^{-1})$ in $l^\infty(\mathcal{X})$, and so

(d) $\sqrt{N}\frac{\widehat{g}^b(x) - \widehat{g}(x)}{\|e(x)\|} =^d \frac{e(x)'}{\|e(x)\|}\mathcal{N}_d + o_P(a_N^{-1})$ in $l^\infty(\mathcal{X})$.

Corollary 2.1 establishes strong approximation of $\alpha(x)'(\widehat{\beta} - \beta_0)$ by a Gaussian process. Theorem 4.5 in Belloni et al. (2015) implies validity of weighted bootstrap.

3 Applications

In this section we apply the results of Section 2 for economically relevant settings, described in Examples 1, 2, 4.

3.1 Conditional Average Treatment Effect

Using the setup of Example 1, let Y^1 and Y^0 be the potential outcomes, $D \in \{1, 0\}$ indicate the presence in the treatment group, $Y^o = DY^1 + (1 - D)Y^0$ be the actual outcome, $s_0(Z) = \mathbb{E}[D = 1|Z]$ be the propensity score and $\mu_0(D, Z) = \mathbb{E}[Y^o|D, Z]$ be the conditional expectation function. We provide sufficient low-level conditions on the regression functions $\mu_0(1, Z), \mu_0(0, Z)$

and the propensity score $s_0(Z)$ such that the pointwise and uniform Gaussian approximations of the target function $g(x)$ (Theorems 2.1 and 2.2) hold.

Assumption 3.1 (Strong Overlap). *A The propensity score is bounded above and below:* $\exists \bar{s}_0 > 0 \quad 0 < \bar{s}_0 < s_0(z) < 1 - \bar{s}_0 < 1 \quad \forall z \in \mathcal{Z}$.

B The propensity score is bounded below: $\exists \bar{s}_0 > 0 \quad 0 < \bar{s}_0 < s_0(z) < 1 \quad \forall z \in \mathcal{Z}$.

In context of Example 1 Assumption 3.1(a) ensures that the probability of assignment to the treatment and control group is bounded away from zero. In context of Example 2 Assumption 3.1(b) ensures that the probability of observing the response Y^* is bounded away from zero.

Definition 3.1 (First Stage Rate). *Given the true functions s_0, μ_0 and sequences of shrinking neighborhoods S_N of s_0 and M_N of μ_0, define the following rates:*

$$\boldsymbol{s}_N := \sup_{s \in \mathcal{S}_N} (\mathbb{E}(s(Z) - s_0(Z))^2)^{1/2},$$

$$\boldsymbol{m}_N := \sup_{\mu \in \mathcal{M}_N} (\mathbb{E}(\mu(Z) - \mu_0(Z))^2)^{1/2},$$

where the expectation is taken with respect to Z.

We will refer to $\mathbf{s}_N$ as the propensity score rate and $\mathbf{m}_N$ as the regression function rate.

Assumption 3.2 (Assumptions on the Propensity Score and the Regression Function). *Assume that there exists a sequence of numbers $\epsilon_N = o(1)$ and sequences of neighborhoods S_N of s_0, M_N of μ_0 with rates $\boldsymbol{s}_N, \boldsymbol{m}_N$ such that the first-stage estimate $\{\widehat{s}(z), \widehat{\mu}(1,z), \widehat{\mu}(0,z)\}$(A) or $\{\widehat{s}(z), \widehat{\mu}(z)$(B) belongs to the set $\{S_N, \mathcal{M}_N\}$ w.p. at least $1 - \epsilon_n$ and*

$$\xi_d \boldsymbol{s}_N \boldsymbol{m}_N \sqrt{\log N} = o(1).$$

Finally, assume that there exists $C > 0$ that bounds the functions in M_N uniformly over their domain (A): $\sup_{\mu \in \mathcal{M}_N} \sup_{z \in \mathcal{Z}} \sup_{d \in \{1,0\}} |\mu(d,z)| < C$ *or (B):* $\sup_{\mu \in \mathcal{M}_N} \sup_{z \in \mathcal{Z}} |\mu(z)| < C$.

Plausibility of Assumption 3.2 is discussed in the introduction of the paper. In case the propensity score and regression function can be well-approximated by a logistic (linear) high-dimensional sparse model, Assumption 3.2 holds under a low-level conditions analogous to those in Example 5.

Corollary 3.1 (Gaussian Approximation for Conditional Average Treatment Effect). *Under Assumptions 3.1(A) and 3.2(A), the robust signal, given by Equation 1.3, satisfies Assumptions 2.5 and 2.9. As a result, pointwise and uniform Gaussian approximation (Theorems 2.1 ,2.2) and Validity of Weighted Bootstrap (Corollary 2.1) hold with*

$$\Omega = Q^{-1} \Sigma Q^{-1},$$

where Σ is:

$$\Sigma = \mathbb{E}p(X_i)p(X_i)'\big[\mu(1,Z_i) - \mu(0,Z_i) + \frac{D_i[Y_i^o - \mu(1,Z_i)]}{s_0(Z_i)} - \frac{(1-D_i)[Y_i^o - \mu(0,Z_i)]}{1-s_0(Z_i)} - p(X_i)'\beta_0\big]^2.$$

3.2 Regression Function with Partially Missing Outcome

Using the setup of Example 2, let Y^* be a partially missing outcome variable, $D \in \{1,0\}$ indicate the presence of Y^* in the sample, $Y^o = DY^*$ be the observed outcome, $s_0(Z) = \mathbb{E}[D=1|Z]$ be the propensity score and $\mu_0(Z) = \mathbb{E}[Y^o|D=1,Z]$ be the conditional expectation function. We provide sufficient low-level conditions on the regression functions $\mu(Z), s(Z)$ such that the pointwise and uniform Gaussian approximations of the target function $g(x)$ (Theorems 2.1 and 2.2) hold.

Corollary 3.2 (Gaussian Approximation for Regression Function with Partially Missing Outcome)**.** *Under Assumptions 3.1(B) and Assumption 3.2(B) the robust signal, given by Equation 1.4, satisfies Assumptions 2.5 and 2.9. As a result, pointwise and uniform Gaussian approximation (Theorems 2.1 ,2.2) and Validity of Weighted Bootstrap (Corollary 2.1) hold with*

$$\Omega = Q^{-1}\Sigma Q^{-1},$$

where Σ is:

$$\Sigma = \mathbb{E}p(X_i)p(X_i)'\big[\frac{D_i}{s_0(Z_i)}[Y_i^o - \mu_0(Z_i)] + \mu_0(Z_i) - p(X_i)'\beta_0\big]^2.$$

Here give an example of a model and and a first-stage estimator that satisfy Assumption 3.2.

Example 5 (Partially Missing Outcome with High-Dimensional Sparse Design)**.** Consider the setup of Example 2. Let the observable vector (D, X, DY^*) consist of the covariate vector of interest X and a partially observed variable Y^*, whose presence is indicated by $D \in \{1,0\}$. In addition, suppose there exists an observable vector Z such that Missingnesss at Random (Assumption 1.3) is satisfied conditionally on Z. Let $p_\mu(Z), p_s(Z)$ be high-dimensional basis functions of the vector Z that approximate the conditional expectation functions $\mu_0(z), s_0(z)$ using the linear and logistic links, respectively:

$$\mu_0(z) = p_\mu(Z)'\theta + r_\mu(z) \tag{3.1}$$

$$s_0(z) = L(p_s(Z)'\delta) + r_s(z) := \frac{\exp(p_s(Z)'\delta)}{\exp(p_s(Z)'\delta)+1} + r_s(z) \tag{3.2}$$

where θ, δ are the vectors in $\mathcal{R}^p$ whose dimension p is allowed to be larger than the sample size N, and $r_\mu(z), r_s(z)$ are the misspecification errors of the respective link functions that vanish as described in Assumptions 3.3, 3.4. For each $\gamma \in \{\theta, \delta\}$, denote a support set

$$T_\gamma := \{j : \gamma_j \neq 0, j \in \{1,2,..,p\}\}$$

and its cardinality, which we refer to as sparsity index of γ,

$$s_\gamma := |T| = \|\gamma\|_0 \quad \forall \quad \gamma \in \{\theta, \delta\}.$$

We allow the cardinality of s_δ, s_θ to grow with N. Define minimal and maximal empirical Restricted Sparse Eigenvalues (RSE) $\phi_{\min}(m), \phi_{\max}(m)$ as

$$\phi_{\min}(m) := \min_{1 \leqslant \|\nu\|_0 \leqslant m} \frac{\nu' \mathbb{E}_N Z_i Z_i' \nu}{\|\nu\|_2^2}, \quad \phi_{\max}(m) := \max_{1 \leqslant \|\nu\|_0 \leqslant m} \frac{\nu' \mathbb{E}_N Z_i Z_i' \nu}{\|\nu\|_2^2}.$$

Let $\delta_N \to 0$ and $\Delta_N \to 0$ be the fixed constants approaching zero from above at a speed at most polynomial in N: for example, $\delta_N \geqslant \frac{1}{N^c}$ for some $c > 0$, $\ell_N = \log N$, and c, C, κ', κ'' and $\nu \in [0, 1]$ are positive constants.

Assumption 3.3 (Regularity Conditions for Linear Link)**.** *We assume that the following standard conditions hold. With probability $1 - \Delta_N$, the minimal and maximal empirical RSE are bounded from below by κ'_μ and from above by κ''_μ:*

$$\kappa'_\mu \leqslant \inf_{\|\delta\|_0 \leqslant s\ell_N, \|\delta\|=1} \|Dp_\mu(Z)\|_{P_N,2} \leqslant \sup_{\|\delta\|_0 \leqslant s\ell_N, \|\delta\|=1} \|Dp_\mu(Z)\|_{P_N,2} \leqslant \kappa''_\mu.$$

(b) There exists absolute constants $B, c > 0$: regressors $\max_{1 \leqslant j \leqslant p} |p_{\mu,j}(Z)| \leqslant B$ a.s. and $\max_{1 \leqslant j \leqslant p} c \leqslant \mathbb{E} p_{\mu,j}(Z)^2$ (c) With probability $1 - \Delta_N$, $\mathbb{E}_N r_\mu^2(Z) \leqslant Cs \log(p \vee N)/N$.

Assumption 3.4 (Regularity Conditions for Logistic Link)**.** *We assume that the following standard conditions hold. With probability $1 - \Delta_N$, the minimal and maximal empirical RSE are bounded from below by κ'_s and from above by κ''_s:*

$$\kappa'_s \leqslant \inf_{\|\delta\|_0 \leqslant s\ell_N, \|\delta\|=1} \|p_s(Z)'\|_{P_N,2} \leqslant \sup_{\|\delta\|_0 \leqslant s\ell_N, \|\delta\|=1} \|p_s(Z)'\delta\|_{P_N,2} \leqslant \kappa''_s.$$

(b) There exist absolute constants $B, c > 0$: regressors $\max_{1 \leqslant j \leqslant p} |p_{s,j}(Z)| \leqslant B$ a.s. and $\max_{1 \leqslant j \leqslant p} c \leqslant \mathbb{E} p_{s,j}(Z)^2$ (c) With probability $1 - \Delta_N$, $\mathbb{E}_N r_s^2(Z) \leqslant Cs \log(p \vee N)/N$.

Assumptions 3.3 and 3.4 are a simplification of the Assumption 6.1-6.2 in Belloni et al. (2013). The following estimators of $\mu_0(Z)$ and $s_0(Z)$ are available.

Definition 3.2 (Lasso Estimator of the Regression Function)**.** *For $\lambda = 1.1\sqrt{N}\Phi^{-1}(1 - 0.05/(N \vee p \log N))$, define $\widehat{\theta}$ as a solution to the following optimization problem:*

$$\widehat{\theta} := \arg\min_{\theta \in \mathcal{R}^p} \mathbb{E}_N D_i (Y_i^o - Z_i'\theta)^2 + \lambda \|\theta\|_1$$

and a first-stage estimate of μ as

$$\widehat{\mu}(z) := z'\widehat{\theta}.$$

Definition 3.3 (Lasso Estimator of the Propensity Score)**.** *For $\lambda = 1.1\sqrt{N}\Phi^{-1}(1 - 0.05/(N \vee p \log N))$ and an appropriately chosen $\underline{s} > 0$, define $\widehat{\delta}$ as a solution to the following optimization*

problem:

$$\widehat{\delta} := \arg\min_{\delta \in \mathcal{R}^p} \mathbb{E}_N[\log(1+\exp(Z_i'\delta)) - D_i Z_i'\delta] + \lambda\|\delta\|_1$$

and a first-stage estimate of s_0 *as*

$$\widehat{s}(z) := \max(\underline{s}/2, L(z'\widehat{\delta})).$$

Lemma 3.1 (Sufficient Conditions for Assumption 2.5)**.** *Suppose Assumptions 3.3 and 3.4 hold. Define the regression rate* $\boldsymbol{m}_N := \sqrt{\frac{s_\theta \log p}{N}}$ *and the propensity score rate* $\boldsymbol{s}_N := \sqrt{\frac{s_\delta \log p}{N}}$*. For a sequence of numbers* $\epsilon_N = o(1)$ *define the sequence of neighborhoods* M_N *of* $\mu_0(Z)$ *and* S_N *of* $s_0(Z)$ *as follows:*

$$M_N := \{\mu(z) = z'\theta : \quad \|\theta - \theta_0\|_{2,N} \leqslant C(-\log(\epsilon_N))\boldsymbol{m}_N\}$$
$$S_N := \{s(z) = \max(\bar{s}/2, L(z'\delta)) : \quad \|\delta - \delta_0\|_{2,N} \leqslant C(-\log(\epsilon_N))\boldsymbol{s}_N\}$$

Then, Assumption 3.2(B) is satisfied if the product of sparsity indices $s_\theta s_\delta$ *grows sufficiently slow:*

$$\sqrt{N}\xi_d \boldsymbol{m}_N \boldsymbol{s}_N \sqrt{\log N} = \xi_d \sqrt{\frac{s_\theta s_\delta \log^2 p \log N}{N}} = o(1).$$

3.3 Conditional Average Partial Derivative

Using the setup of Example 4, let Y^o be an outcome of interest, $\mu(x,w) := \mathbb{E}[Y^o|X=x, W=w]$ be a conditional expectation of Y^o on X, W, and $f(W|X)$ be the conditional density of W given X. We provide sufficient low-level conditions on the regression functions $f(W|X), \mu(X,W)$ such that the pointwise and uniform Gaussian approximations of the target function $g(x)$ (Theorems 2.1 and 2.2) hold.

Definition 3.4 (First Stage Rate)**.** *Given a true function* $f_0(W|X), \mu_0(X,W)$*, let* F_N, M_N *be a sequence of shrinking neighborhoods of* $f_0(W|X)$ *and* $\mu_0(X,W)$ *constrained as follows:*

$$\boldsymbol{f}_N := \sup_{f \in \mathcal{F}_N} (\mathbb{E}(f(W|X) - f_0(W|X))^2)^{1/2}$$

$$\boldsymbol{m}_N := \sup_{\mu \in \mathcal{M}_N} (\mathbb{E}(\mu(X,W) - \mu_0(X,W))^2)^{1/2}$$

where expectation is taken with respect to W, X*.*

We will refer to $\mathbf{f}_N$ as the density rate and $\mathbf{m}_N$ as regression function rate.

Assumption 3.5 (Assumptions on the Conditional Density and the Regression Function)**.** *Assume that there exists a sequence of numbers* $\epsilon_N = o(1)$ *and a sequence of neighborhoods* F_N, M_N

such that the first-stage estimate $\{\widehat{f}, \widehat{\mu}\}$ *belongs to the set* $\{F_N, \mathcal{M}_N\}$ *w.p. at least* $1 - \epsilon_n$*. The neighborhoods* F_N, M_N *shrink at rates* $\boldsymbol{f}_N, \boldsymbol{m}_N$ *such that:*

$$\xi_d \boldsymbol{f}_N \boldsymbol{m}_N \sqrt{\log N} = o(1).$$

Finally, assume that there exists $C > 0$ *that bounds the functions in* F_N, M_N *uniformly over their domain:*

$$\sup_{\mu \in \mathcal{M}_N} \sup_{x,w \in \mathcal{X} \times \mathcal{W}} |\mu(x,w)| < C$$

and

$$\sup_{f \in \mathcal{F}_N} \sup_{x,w \in \mathcal{X} \times \mathcal{W}} |f(W = w | X = x)| < C.$$

Corollary 3.3 (Gaussian Approximation for Conditional Average Partial Derivative)**.** *Let Assumption 3.5 hold. Then, the robust signal, given by Equation 1.4, satisfies Assumptions 2.5 and 2.9. As a result, pointwise and uniform Gaussian approximation (Theorems 2.1 ,2.2) and Validity of Weighted Bootstrap (Corollary 2.1) hold with*

$$\Omega = Q^{-1} \Sigma Q^{-1},$$

where Σ *is:*

$$\Sigma = \mathbb{E} p(X_i) p(X_i)' \big[- \partial_w \log f(W_i | X_i)[Y_i^o - \mu_0(X_i, W_i)] + \mu_0(X_i, W_i) - p(X_i)' \beta_0 \big]^2.$$

The regression function $\mu(X, W)$ can be estimated at $o(N^{-1/4})$ rate by a local linear estimator with suitably chosen bandwdith parameters. An example of a conditional density $f(W|X)$ estimate is a kernel density estimator.

4 Simulation Evidence

In this section we examine the finite sample performance of the Locally Robust Estimator through the Monte Carlo experiments in context of Example 5. We compare LRE to other more naive strategies, such as Inverse Probability Weighting (IPW) and Ordinary Least Squares (OLS) on the complete data only. We show that under the small misspecification of a linear model, all three estimators have similar performance, while under larger misspecification, only Locally Robust Estimator remains valid.

Let us describe our simulation design. Using the setup of Example 5, we generate a random sample $(D_i, X_i, Z_i, Y_i^*)_{i=1}^{N=500}$ from the following data generating process. The control vector Z, $\dim(Z) = 500$, $Z \sim N(0, T(\rho)), \rho = 0.5$ is generated from a normal distribution $N(0, T(\rho))$, where the covariance matrix $T(\rho)$ is the Toeplitz covariance matrix with the correlation parameter $\rho = 0.5$. The propensity score $s_0(z) = L(z'\delta)$, where $L(t) := \dfrac{\exp t}{\exp t + 1}$ is the logistic function, and the parameter $\delta = (1, \frac{1}{2}, \ldots, \frac{1}{100}, 0, \ldots, 0)$. The regression function $\mu(z) = z'\gamma$ is a linear function

of the control vector, where the parameter $\gamma = [(1, \frac{1}{2^2}, \ldots, \frac{1}{(d-1)^2})', c(\frac{1}{d^2}, \ldots, \frac{1}{300}^2)', 0, \ldots, 0]$ and c is a design constant. The outcome variable Y and the presence indicator D are generated by

$$\begin{aligned} D &\sim B(L(Z'\delta)), \\ Y^* &= Z'\theta + \epsilon, \quad \epsilon \sim N(0,1), \end{aligned} \tag{4.1}$$

where $B(p)$ stands for a Bernoulli draw with probability of success p. Suppose a researcher is interested in the conditional expectation of Y given the first $d = 6$ control variables $X = [Z_1, Z_2, ..., Z_d]$:

$$g(x) := \mathbb{E}[Y^*|X = x].$$

He approximates $g(x)$ using a linear form $p(x)'\beta$, where the vector of technical transformations

$$p(x) := (1, x)'$$

[4] consists of the constant and a degree one polynomial of vector x. Let $Y^o = DY^*$ be the observed outcome. Having established the setup, let us describe the estimators whose finite sample performance we compare:

- Ordinary Least Squares: $\widehat{\beta}_{OLS} := (\mathbb{E}_N D_i p(X_i) p(X_i)')^{-1} \mathbb{E}_N D_i p(X_i) Y_i^o$
- Inverse Probability Weighting: $\widehat{\beta}_{IPW} := \left(\mathbb{E}_N \frac{D_i}{\widehat{s}(Z_i)} p(X_i) p(X_i)'\right)^{-1} \mathbb{E}_N \frac{D_i}{\widehat{s}(Z_i)} p(X_i) Y_i^o$
- Locally Robust Estimator: $\widehat{\beta}_{LRE} := (\mathbb{E}_N p(X_i) p(X_i)')^{-1} \mathbb{E}_N p(X_i) \big[\frac{D_i}{\widehat{s}(Z_i)} [Y_i^o - \widehat{\mu}(Z_i)] + \widehat{\mu}(Z_i)\big]$

where the nonparametric estimates of the propensity score $\widehat{s}$ and the regression function $\widehat{\mu}$ are estimated as in Example 5 using the cross-fitting procedure in Definition 1.1.

Table 1 shows the bias, standard deviation, root mean squared error, and rejection frequency for Ordinary Least Squares, Inverse Probability Weighting, and Locally Robust Estimator under small misspecification, which is achieved by scaling the coefficient on the omitted controls by a small constant ($c = 0.1$). In that case, all the three estimators have small bias and good coverage property. Since the linear model is close to the true one, OLS is best linear conditionally unbiased estimator, and therefore has smaller variance than IPW and LRE.

Table 2 shows finite sample properties of IPW, LRE, and OLS under large misspecification, which is achieved by scaling the coefficient on the omitted controls by a small constant ($c = 20$). As expected, OLS suffers from selection bias, IPW incurs the first-order bias due to the propensity score estimation error, but LRE remains valid. In the case of large misspecification, LRE achieves 8% to 100% bias reduction compared to IPW. Moreover, LRE maintains valid inference and has its rejection frequency close to the nominal under both small and large misspecification.

[4] We omit the regressor $p(x) = 1$ since $\mathbb{E}Y^* = \mathbb{E}Z = 0$ by construction.

	OLS	IPW	LRE	OLS	IPW	LRE	OLS	IPW	LRE	OLS	IPW	LRE
		Bias			St.Error			RMSE			Rej.Freq.	
$\beta_1 = 1\ 1$	0.001	-0.002	-0.014	0.049	0.060	0.061	0.049	0.060	0.063	0.080	0.080	0.090
$\beta_2 = 0.5$	0.005	0.003	0.004	0.101	0.126	0.124	0.102	0.126	0.125	0.067	0.090	0.100
$\beta_3 = 0.3$	-0.010	-0.004	-0.009	0.091	0.120	0.119	0.092	0.120	0.119	0.030	0.037	0.050
$\beta_4 = 0.2$	-0.000	-0.002	-0.002	0.096	0.118	0.116	0.096	0.118	0.116	0.037	0.060	0.050
$\beta_5 = 0.2$	0.002	0.005	0.004	0.100	0.116	0.118	0.100	0.116	0.118	0.070	0.060	0.057

Table 1: Bias, St.Error, RMSE, Rejection Frequency of OLS, IPW, LRE. Small misspecification. Test size $\alpha = 0.05$, design constant $c = 0.1$, $R^2 = 0.5$, Number of Monte Carlo Repetitions 300.

	OLS	IPW	LRE	OLS	IPW	LRE	OLS	IPW	LRE	OLS	IPW	LRE
		Bias			St.Error			RMSE			Rej.Freq	
$\beta_1 = 1\ 1$	0.109	0.113	0.023	0.863	1.156	0.655	0.869	1.161	0.656	0.097	0.107	0.110
$\beta_2 = 0.5$	-0.233	-0.456	-0.092	1.806	2.052	1.346	1.821	2.102	1.349	0.083	0.107	0.120
$\beta_3 = 0.3$	0.025	0.127	-0.056	1.916	2.200	1.339	1.916	2.203	1.340	0.120	0.100	0.103
$\beta_4 = 0.2$	-0.060	-0.007	-0.004	1.886	2.284	1.353	1.887	2.284	1.353	0.093	0.143	0.113
$\beta_5 = 0.2$	0.065	0.001	-0.019	1.865	2.274	1.303	1.866	2.274	1.303	0.113	0.117	0.087

Table 2: Bias, St.Error, RMSE, Rejection Frequency of OLS, IPW, LRE. Large misspecification. Test size $\alpha = 0.1$, design constant $c = 20$, $R^2 = 0.2$, Number of Monte Carlo Repetitions 300.

5 Empirical Application

We apply our methods to study the household demand for gasoline, a question studied in Hausman and Newey (1995), Schmalensee and Stoker (1999), Yatchew and No (2001) and Blundell et al. (2012). These papers estimated the demand function and the average price elasticity for various demographic groups. The dependence of the price elasticity on the household income was highlighted in Blundell et al. (2012), who have estimated the elasticity by low, middle, and high-income groups and found its relationship with income to be non-monotonic. To gain more insight into this question, we estimate the average price elasticity as a function of income and provide simultaneous confidence bands for it.

The data for our analysis are the same as in Yatchew and No (2001), coming from National Private Vehicle Use Survey, conducted by Statistics Canada between October 1994 and September 1996. The data set is based on fuel purchase diaries and contains detailed information about fuel prices, fuel consumption patterns, vehicles and demographic characteristics. We employ the same selection procedure as in Yatchew and No (2001) and Belloni et al. (2011), focusing on a sample of the households with non-zero licensed drivers, vehicles, and distance driven which leaves us with 5001 observations.

The object of interest is the average predicted percentage change in the demand due to a unit percentage change in the price, holding the observed demographic characteristics fixed, conditional on income. In context of Example 4, this corresponds to the conditional average derivative

$$g(x) = \mathbb{E}[\partial_w \mu(X, Z, W)|X = x],$$
$$\mu(w, x, z) = \mathbb{E}[Y^o|X = x, Z = z, W = w],$$

where Y^o is the logarithm of gas consumption, W is the logarithm of price per liter, X is log income, and Z are the observed subject characteristics such as household size and composition,

distance driven, and the type of fuel usage. The robust signal Y for the target function $g(x)$ is given by

$$Y = -\partial_w \log f(W|X,Z)(Y^o - \mu(X,Z,W)) + \partial_w \mu(X,Z,W), \tag{5.1}$$

where $f(w|x,z) = f(W = w|X = x, Z = z)$ is the conditional density of the price variable W given income X and subject characteristics Z. The conditional density $f(w|x,z)$ and the conditional expectation functions $\mu(w,x,z)$ comprise the set of the nuisance parameters to be estimated in the first stage.

The choice of the estimators in the first and the second stages is as follows. To estimate the conditional expectation function $\mu(w,x,z)$ and its partial derivative $\partial_w \mu(w,x,z)$, we consider a linear model that includes price, price squared, income, income squared, their interactions with 28 time, geographical, and household composition dummies. All in all, we have 91 explanatory variables. We estimate $\mu(w,x,z)$ using Lasso with the penalty level chosen as in Belloni et al. (2014), and estimate the derivative $\partial_w \mu(w,x,z)$ using the estimated coefficients of $\mu(w,x,z)$. To estimate the conditional density $f(w|x,z)$, we consider a model:

$$W = l(X,Z) + U, \quad U \perp X, Z,$$

where $l(x,z) = \mathbb{E}[W|X = x, Z = z]$ is the conditional expectation of price variable W given income variable X and covariates Z, and U is an independent continuously distributed shock with univariate density $\phi(\cdot)$. Under this assumption, the log density $\partial_w \log f(W|X,Z)$ equals to

$$\partial_w \log f(W = w|X = x, Z = z) = \frac{\phi'(w - l(x,z))}{\phi(w - l(x,z))}.$$

We estimate $\phi(u) : \mathcal{R} \to \mathcal{R}^+$ by an adaptive kernel density estimator of Portnoy and Koenker (1989) with Silverman choice of bandwidth. Finally, we plug in the estimates of $\mu(w,x,z)$, $\partial_w \mu(w,x,z)$, $f(w|z,x)$ into the Equation 5.1 to get an estimate of the robust signal $\widehat{Y}$ and estimate $g(x)$ by least squares series regression of $\widehat{Y}$ on X. We try both polynomial basis function and B-splines to construct technical regressors.

Figures 1 and 3 report the estimate of the target function (the black line), the pointwise (the dashed blue lines) and the uniform confidence (the solid blue lines) bands for the average price elasticity conditional on income, where the significance level $\alpha = 0.05$. The uniform confidence bands for $g(x)$ are chosen such that they contain the true function $g(x)$ with probability $1 - \alpha = 0.95$. The are computed as:

$$[\widehat{g}(x) - \widehat{e}(x) t^*_{1-\alpha}, \widehat{g}(x) + \widehat{e}(x) t^*_{1-\alpha}],$$

where $\widehat{g}(x)$ is the estimate of the target function, $\widehat{e}(x) = \sqrt{p(x)'\widehat{\Omega} p(x)}$ is the estimate of the standard error, $\widehat{\Omega}$ is the estimated asymptotic variance of $\widehat{\beta}$ (Eq. 2.5), and the $t^*_{1-\alpha}$-statistic is

the $(1-\alpha)$- empirical quantile of the t-statistic bootstrap distribution

$$t^b := \sup_{x \in \mathcal{X}} |\frac{p(x)'(\widehat{\beta}^b - \widehat{\beta})}{\widehat{e}(x)}|.$$

The panels of Figure 1 correspond to different choices of the first-stage estimates of the nuisance functions $\mu(w, x, z)$ and $f(w|x, z)$ and dictionaries of technical regressors. The panels of Figure 3 correspond to the subsamples of large and small households and to different choices of the dictionaries.

The summary of our empirical findings based on Figure 1 and 3 is as follows. We find the elasticity to be in the range $(-1, 0)$ and significant for majority of income levels. The estimates based on B-splines (Figures 1c, 1d) are monotonically increasing in income, which is intuitive. The estimates based on polynomial functions are non-monotonic in income. For every algorithm on Figure 1 we cannot reject the null hypothesis of constant price elasticity for all income levels: for each estimation procedure, the uniform confidence bands contain the constant function. Figure 3 shows the average price elasticity conditional on income for small and large households.[5]. For majority of income levels, we find large households to be more price elastic than the small ones, but the difference is not significant at any income level.

To demonstrate the relevance of demographic data Z in the first stage estimation, we have shown the average predicted effect of the price change on the gasoline consumption (in logs), without accounting for the covariates in the first stage. In particular, this effect equals to $\mathbb{E}[\partial_w \mu(X, W)|X = x]$, where $\mu(x, w) = \mathbb{E}[Y|X = x, W = w]$ is the conditional expectation of gas consumption given income and price. This predictive effect consists two effects: the effect of price change on the consumption holding the demographic covariates fixed, which we refer to as average price elasticity, and the association of the price change with the change in the household characteristics that also affect the consumption themselves. Figure 2 shows this predictive effect, approximated by the polynomials of degree $k \in \{1, 2\}$, conditional on income. By contrast to the results in Figure 1, the slope of the polynomial of degree $k = 1$ has a negative relationship between income and price elasticity, which present evidence that the demographics Z confound the relationship between income and price elasticity.

6 Proofs

6.1 Notation

We will use the following notation. Let

$$\mathbb{E}_{n,k} f(x_i) := \frac{1}{n} \sum_{i \in J_k} f(x_i), \quad \mathbb{G}_{n,k} f(x_i) := \frac{1}{n} \sum_{i \in J_k} f(x_i) - \mathbb{E}\big[f(x_i)|J_k^c\big]$$

[5] A large household is a household with at least 4 members.

For an observation index $i \in J_k$ that belongs to a fold $J_k, k \in \{1, 2, .., K\}$, define $Y_i(\widehat{\eta}) = Y_i(\widehat{\eta}_k), i \in J_k$, where $\widehat{\eta}_k$ is estimated on J_k^c as in Definition 1.1. Define $\widehat{\eta}(Z_i) = \widehat{\eta}_k(Z_i), i \in J_k$.

Let $p(x_i) := p_i$, $\widehat{Q} := \mathbb{E}_N p_i p_i'$, and $r_i := r_g(x_i) = g(x_i) - p_i'\beta_0$. For two sequences of random variables denote $a_N, b_N, n \geqslant 1 : a_N \lesssim_P b_n$ means $a_N = O_P(b_N)$. For two sequences of numbers, denote $a_N, b_N, n \geqslant 1 : a_N \lesssim b_n := a_N = O(b_N)$. Let $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$. The l_2 norm is denoted by $\|\cdot\|$, the l_1 norm is denoted by $\|\cdot\|_1$, the l_∞ is denoted by $\|\cdot\|_\infty$, and the l_{i0} - norm denotes the number of nonzero components of a vector.

Given a vector $\delta \in \mathcal{R}^p$ and a set of indices $T \subset \{1, ..., p\}$, we denote by δ_T the vector in $\mathcal{R}^p$ in which $\delta_{Tj} = \delta_j, j \in T$ and $\delta_{Tj} = 0, j \notin T$.

6.2 Technical Lemmas

Theorem 6.1 (LLN for Matrices)**.** *Let $Q_i = p_i p_i'^\top$ be i.i.d symmetric non-negative kxk-matrices with $d \geqslant e^2$. Notice that $\|Q_i\| = \|p_i\|^2 \leqslant \xi_d^2$. Let $Q = \frac{1}{N}\sum_{i=1}^N \mathbb{E} p_i p_i'^\top$ denote average value of population covariance matrices.*

$$\mathbb{E}\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}$$

Proof. Proof can be found in Rudelson (1999). ∎

Lemma 6.1 (Conditional Convergence Implies Unconditional)**.** *Let $\{X_m\}_{m\geqslant 1}$ and $\{Y_m\}_{m\geqslant 1}$ be sequences of random vectors. (i) If for $\epsilon_m \to 0, \mathrm{P}(\|X_m\| > \epsilon_m | Y_m) \to_P 0$, then $\mathrm{P}(\|X_m\| > \epsilon_m) \to 0$. In particular, this occurs if $\mathbb{E}[\|X_m\|^q/\epsilon_m^q | Y_m] \to_P 0$ for some $q \geqslant 1$, by Markov inequality. (ii) Let $\{A_m\}_{m\geqslant 1}$ be a sequence of positive constants. If $\|X_m\| = O_P(A_m)$ conditional on Y_m, namely, that for any $\ell_m \to \infty, \mathrm{P}(\|X_m\| > \ell_m A_m | Y_m) \to_P 0$, then $X_m = O_P(A_m)$ unconditionally, namely, that for any $\ell_m \to \infty, \mathrm{P}(\|X_m\| > \ell_m A_m) \to 0$.*

Proof. The Lemma is a restatement of Lemma 6.1 of Chernozhukov et al. (2016a) ∎

Lemma 6.2 (Maximal Inequality)**.** *Let an i.i.d sample of size n be available. Let $\mathcal{F}$ be a function class with an envelope $F \geqslant \sup_{f\in\mathcal{F}} |f|$ with $\|F\|_{P,q} < \infty$ for some $q \geqslant 2$ Let $M := \max_{i\leqslant n} F(W_i)$ and $\sigma^2 > 0$ be any positive constant such that*

$$\sup_{f\in\mathcal{F}} \|f\|_{P,2}^2 \leqslant \sigma^2 \leqslant \|F\|_{P,2}^2$$

Suppose there exist constants $a \geqslant e$ and $v \geqslant 1$ such that

$$\log \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant v \log(a/\epsilon), 0 < \epsilon \leqslant 1$$

Then,

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

where $\|M\|_{P,q} \leqslant n^{1/q}\|F\|_{P,s}$ *Moreover, with probability at least* $1 - c(\log n)^{-q/2}$,

$$\sup_{f\in\mathcal{F}}\|\mathbb{G}_n f\| \leqslant K(q,c)\left(\sigma\sqrt{v\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,q}}{\sqrt{n}}\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)\right)$$

Proof. The Lemma is a restatement of Lemma 6.2 Maximal Inequality of Chernozhukov et al. (2016a). ∎

Lemma 6.3 (No Effect of First Stage error)**.**

$$\sqrt{N}\|\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]\| = O_P(B_N + \xi_d\kappa_N) = o(1) \tag{6.1}$$

Proof. Define an event $\mathcal{E}_N := \{\widehat{\eta}_k \in T_N \quad \forall k \in [K]\}$, such that the nuisance parameter estimate $\widehat{\eta}_k$ belongs to the realization set T_N for each fold $k \in [K]$. By union bound, this event holds w.h.p.

$$\mathrm{P}_{P_N}(\mathcal{E}_N) \geqslant 1 - K\epsilon_N = 1 - o(1).$$

$$\begin{aligned}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)] = \frac{1}{K}\sum_{k=1}^{K}&\underbrace{\mathbb{E}_{n,k}p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)] - \mathbb{E}\big[p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]|(W_i)_{i\in J_k^c}\big]}_{I_{1,k}}\\ &+\underbrace{\mathbb{E}\big[p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)|(W_i)_{i\in J_k^c}\big] - \mathbb{E}p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]}_{I_{2,k}}\end{aligned}$$

Conditionally on $(W_i)_{i\in J_k}$, the estimator $\widehat{\eta} = \widehat{\eta}_k$ is non-stochastic. On the event $\mathcal{E}_N$

$$\begin{aligned}\mathbb{E}[\|\sqrt{n}I_{1,k}\|^2|\mathcal{E}_N,(W_i)_{i\in J_k^c}] &\leqslant \mathbb{E}[\|p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]\|^2|\mathcal{E}_N,(W_i)_{i\in J_k^c}]\\ &\leqslant \sup_{\eta\in T_N}\mathbb{E}\|p_i[Y_i(\eta) - Y_i(\eta_0)]\|^2\\ &\leqslant \sup_{x\in\mathcal{X}}\|p(x)\|^2\sup_{\eta\in T_N}\mathbb{E}(Y_i(\eta) - Y_i(\eta_0))^2\\ &\leqslant \xi_d^2\kappa_N^2 \qquad \text{(Assumption 2.5)}\end{aligned}$$

Hence, $\sqrt{n}I_{1,k} = O_{P_N}(\xi_d\kappa_N)$ by Lemma 6.1. To bound $I_{2,k}$, recognize that on the event $\mathcal{E}_N$

$$\begin{aligned}\mathbb{E}[\|\sqrt{n}I_{2,k}\||\mathcal{E}_N,(W_i)_{i\in J_k^c}] &\leqslant \sup_{\eta\in T_N}\sqrt{n}\|\mathbb{E}p_i(Y_i(\eta) - Y_i(\eta_0))|(W_i)_{i\in J_k}\|\\ &\leqslant \sup_{\eta\in T_N}\sqrt{n}\|\mathbb{E}p_i(Y_i(\eta) - Y_i(\eta_0))\| \leqslant B_N\end{aligned}$$

Therefore, $\sqrt{n}I_{2,k} = O_{P_N}(B_N)$ and $\sum_{k=1}^{K}\frac{1}{K}\sqrt{N}(I_{1,k} + I_{2,k}) = O_{P_N}(\xi_d\kappa_N + B_N)$ ∎

Lemma 6.4 (Comparison of Conditional Variance)**.** *If $X \subseteq Z$, the doubly robust signal Y of Example 2 achieves lower conditional variance:*

$$Var(Y|X=x) \leqslant Var(Y^w|X=x) \quad \forall x \in \mathcal{X}$$

than the naive signal Y^w.

Proof. Consider the setup of Example 2:

$$\begin{aligned}
\mathbb{E}[Y^2|Z] &= \mathbb{E}[(Y^w)^2|Z] + 2\mathbb{E}[Y^* \frac{D}{s_0(Z)}(1-\frac{D}{s_0(Z)})\mu_0(Z)|Z] \\
&+ \mathbb{E}[(1-\frac{D}{s_0(Z)})^2\mu_0^2(Z)|Z] \\
&= \mathbb{E}[(Y^w)^2|Z] + \mu_0^2(Z)\mathbb{E}[2\frac{D}{s_0(Z)}(1-\frac{D}{s_0(Z)}) + (1-\frac{D}{s_0(Z)})^2]|Z \\
&= \mathbb{E}[(Y^w)^2|Z] + \mu_0^2(Z)\mathbb{E}[1-\frac{D}{s_0^2(Z)}]|Z \\
&= \mathbb{E}[(Y^w)^2|Z] + \mu_0^2(Z)(1-\frac{1}{s_0(Z)}) \leqslant \mathbb{E}[(Y^w)^2|Z]
\end{aligned}$$

Therefore, $\mathbb{E}[Y^2|X] \leqslant \mathbb{E}[(Y^w)^2|X]$. Since the signals Y, Y^w are unbiased:

$$\mathbb{E}[Y|X=x] = \mathbb{E}[Y^w|X=x] = g(x),$$

$$Var(Y|X) \leqslant Var(Y^w|X).$$

■

Lemma 6.5 (Comparison of Conditional Variance)**.** *If $X \subseteq Z$, the doubly robust signal Y of Example 1 achieves lower conditional variance:*

$$Var(Y|X=x) \leqslant Var(Y^w|X=x) \quad \forall x \in \mathcal{X}$$

than the naive signal Y^w.

Proof. Consider the setup of Example 1: Using the proof of Lemma 6.4, we conclude that

$$\mathbb{E}(\frac{DY^1}{s_0(Z)} + \mu_0(1,Z)(1-\frac{D}{s_0(Z)}))^2|Z \leqslant \mathbb{E}(\frac{DY^1}{s_0(Z)})^2|Z \tag{6.2}$$

$$\mathbb{E}(\frac{(1-D)Y^0}{1-s_0(Z)} + \mu_0(0,Z)(1-\frac{1-D}{1-s_0(Z)}))^2|Z \leqslant \mathbb{E}(\frac{(1-D)Y^0}{1-s_0(Z)})^2|Z \tag{6.3}$$

Therefore,

$$
\begin{aligned}
\mathbb{E}(Y^w)^2|Z &= \mathbb{E}(\frac{DY^1}{s_0(Z)})^2|Z + \mathbb{E}(\frac{(1-D)Y^0}{1-s_0(Z)})^2|Z && (D(1-D)=0)\\
&\geqslant \mathbb{E}(\frac{DY^1}{s_0(Z)} + \mu_0(1,Z)(1-\frac{D}{s_0(Z)}))^2|Z + \mathbb{E}(\frac{(1-D)Y^0}{1-s_0(Z)} + \mu_0(0,Z)(1-\frac{1-D}{1-s_0(Z)}))^2|Z \\
& && ((6.2),\ (6.3))\\
&= \mathbb{E}(\frac{DY^1}{s_0(Z)} + \mu_0(1,Z)(1-\frac{D}{s_0(Z)}))^2|Z - \frac{(1-D)Y^0}{1-s_0(Z)} - \mu_0(0,Z)(1-\frac{1-D}{1-s_0(Z)}))^2|Z\\
&= \mathbb{E}Y^2|Z
\end{aligned}
$$

Therefore, $\mathbb{E}[Y^2|X] \leqslant \mathbb{E}[(Y^w)^2|X]$. Since the signals Y, Y^w are unbiased: $\mathbb{E}[Y|X=x] = \mathbb{E}[Y^w|X=x] = g(x)$,

$$Var(Y|X) \leqslant Var(Y^w|X).$$

■

6.3 Proofs of Theorems

Proof of Theorem 2.1 (a).

$$
\begin{aligned}
\|\widehat{\beta} - \beta_0\| &= \|\widehat{Q}^{-1}\mathbb{E}_N p_i Y_i(\widehat{\eta}) - \beta_0\|\\
&\leqslant \underbrace{\|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]\|}_{S_1} + \underbrace{\|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i[Y_i(\eta_0) - p_i'\beta_0]\|}_{S_2}\\
&= S_1 + \|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i \underbrace{[Y_i(\eta_0) - g(x_i)]}_{u_i}\| + \|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i \underbrace{[g(x_i) - p_i'\beta_0]}_{r_i}\|\\
&= S_1 + \|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i u_i\| + \|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i r_i\|
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbb{E}_N p_i u_i\| &\lesssim_P (\mathbb{E}\big[\|\mathbb{E}_n p_i u_i\|^2\big])^{1/2} && \text{(Markov)}\\
&\leqslant (\mathbb{E}u_i^2 p_i' p_i/N)^{1/2} && \text{(i.i.d data)}\\
&\leqslant \bar{\sigma}\sqrt{d/N} && (6.4)
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbb{E}_N p_i r_i\| &\lesssim_P (\mathbb{E}\big[\|\mathbb{E}_n p_i r_i\|^2\big])^{1/2} && \text{(Markov)}\\
&\leqslant l_d r_d \sqrt{\frac{\mathbb{E}\|p_i\|^2}{N}} = l_d r_d \sqrt{\frac{d}{N}} && \text{(Assumption 2.3)}
\end{aligned}
$$

Alternatively,

$$
\begin{aligned}
\|\mathbb{E}_N p_i r_i\| &\lesssim_P (\mathbb{E}\big[\|\mathbb{E}_n p_i r_i\|^2\big])^{1/2} && \text{(Markov)}\\
&\leqslant \xi_d \sqrt{\frac{\mathbb{E}r_i^2}{N}} = \xi_d r_d/\sqrt{N} && \text{(Assumption 2.3)}
\end{aligned}
$$

With high probability, $\|\widehat{Q}^{-1}\| \leqslant 2\|Q^{-1}\| \leqslant 2/\lambda_{\min}$. Lemma 6.3 implies

$$\|S_1\| \leqslant \|\widehat{Q}^{-1}\|\|\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)\| \leqslant \|\widehat{Q}^{-1}\|\big[B_N/\sqrt{N} + \xi_d\kappa_N/\sqrt{N}\big] = o(\frac{1}{\sqrt{N}})$$

■

Proof of Theorem 2.1 (b). By Definition 1.2,

$$\widehat{\beta} = \widehat{Q}^{-1}\mathbb{E}_N p_i Y_i(\widehat{\eta})$$

Decomposing

$$\begin{aligned}
\sqrt{N}[\mathbb{E}_N p_i Y_i(\widehat{\eta}) - \widehat{Q}\beta_0] &= \sqrt{N}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)] + \sqrt{N}\mathbb{E}_N p_i[Y_i(\eta_0) - p_i'\beta_0] \\
&= \sqrt{N}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)] + \mathbb{G}_N p_i[Y_i(\eta_0) - p_i'\beta_0] \\
&= \sqrt{N}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)] + \mathbb{G}_N p_i u_i + \mathbb{G}_N p_i r_i
\end{aligned}$$

we obtain:

$$\begin{aligned}
\sqrt{N}\alpha'[\widehat{\beta} - \beta_0] &= \sqrt{N}\alpha'\widehat{Q}^{-1}[\mathbb{E}_N p_i Y_i(\widehat{\eta}) - \widehat{Q}\beta_0] && (6.5) \\
&= \sqrt{N}\alpha'\widehat{Q}^{-1}[\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]] && (6.6) \\
&+ \alpha'\widehat{Q}^{-1}\mathbb{G}_N[p_i[r_i + u_i]] \\
&= \alpha' Q^{-1}\mathbb{G}_N[p_i[r_i + u_i]] \\
&+ \underbrace{\alpha^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{G}_N[p_i(u_i + r_i)]}_{I_1} \\
&+ \underbrace{\sqrt{N}\alpha^\top Q^{-1}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]}_{I_2} \\
&+ \underbrace{\sqrt{N}\alpha^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]}_{I_3}
\end{aligned}$$

Total remainder term equals:

$$R_{1,N}(\alpha) = I_1 + I_2 + I_3$$

Decomposing I_1 into sampling and approximation parts:

$$I_1 = \underbrace{\sqrt{N}\alpha^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{E}_N p_i u_i}_{I_{1,a}} + \underbrace{\alpha^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{G}_N p_i r_i}_{I_{1,b}}$$

Definition of regression error $\mathbb{E}[u_i|x_i] = 0$ and $\mathbb{E}[u_i^2|x_i] \lesssim \bar{\sigma}^2$ yields:

$$\begin{aligned}
&\mathbb{E}[I_{1,a}|(x_i)_{i=1}^N] = 0 \\
&\mathbb{E}[I_{1,a}^2|(x_i)_{i=1}^N] \leqslant \alpha^\top[\widehat{Q}^{-1} - Q^{-1}]Q[\widehat{Q}^{-1} - Q^{-1}]\alpha\bar{\sigma}^2 \\
&\qquad\qquad \leqslant \frac{\xi_d^2 \log N}{N}\bar{\sigma}^2 \qquad\qquad \text{(Lemma 6.1)}
\end{aligned}$$

Therefore, $I_{1,a} = o_P(\sqrt{\frac{\xi_d^2 \log N}{N}})$. Using similar argument,

$$|I_{1,b}| \leqslant \sqrt{\frac{\xi_d^2 \log N}{N}}[l_d r_d \sqrt{d} \wedge \xi_d r_d]$$

$$\begin{aligned}
&|I_2| \lesssim_P \|Q\|^{-1}\|\sqrt{N}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]\| \lesssim_P 1/\lambda_{\min}[\xi_d\kappa_N + B_N] = o_P(1) \\
&|I_3| \leqslant \|\alpha\|\|[\widehat{Q}^{-1} - Q^{-1}]\|\|\sqrt{N}\mathbb{E}_N p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]\| \\
&\quad \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}[\xi_d\kappa_N + B_N] = o(1)
\end{aligned}$$

Therefore, with probability approaching one,

$$\sup_{\eta \in T_N} \|R_{1,N}(\alpha)\| \leqslant B_N \vee \xi_d\kappa_N \vee \sqrt{\frac{\xi_d^2 \log N}{N}}[1 + l_d r_d \sqrt{d} \wedge \xi_d r_d]$$

■

Proof of Theorem 2.1 (c). Proof of Theorem 2.1 (c) follows from Theorem 4.2 in Belloni et al. (2015). ■

Lemma 6.6 (Lemma 2, Matrix Convergence Theory Hansen (2014))**.** *Suppose $(w_i)_{i=1}^N$ be i.i.d sequence of d-vectors such that $\|w_i\| \leqslant \xi_d^2$ a.s. Suppose that for some $p > 2$*

$$(\mathbb{E}\|w_i\|^p)^{1/p} \leqslant \xi_d$$

$$\ell_N = \begin{cases} N^{-1/2}\xi_d^{p/(p-2)}(\log d)^{(p-4)/(2p-4)}, & \text{if } p > 4 \\ N^{-(1-2/p)}\xi_d^2 d^{4/p-1} & \text{if } 2 < p \leqslant 4 \end{cases}$$

Then, w.p. $\to 1$,

$$\|\mathbb{E}_N w_i w_i' - \mathbb{E} w_i w_i'\| \lesssim_P d_N$$

Proof. Proof of Theorem 2.1 (d)]

Step 1. We will show that $\widehat{\Sigma} := \mathbb{E}_N p_i p_i'[\widehat{Y}_i - p_i\widehat{\beta}]^2$ is consistent for $\Sigma = \mathbb{E} p_i p_i'[Y_i - p_i\beta_0]^2$ in Step 2. Incorporating the consequence of LLN for Matrices (Lemma 6.1)

$$\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}$$

Therefore,

$$\|\widehat{\Omega}-\Omega\| = \|\widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1} - Q^{-1}\Sigma Q^{-1}\| \lesssim_P o(1)$$

Step 2.

$$\begin{aligned}\|\widehat{\Sigma}-\Sigma\| &= \underbrace{\|\mathbb{E}_N p_i p_i'[\widehat{Y}_i - Y_i]^2\|}_{K_1} + \underbrace{\|\mathbb{E}_N p_i p_i'[Y_i - p_i'\beta_0]^2 - \mathbb{E} p_i p_i'[Y_i - p_i'\beta_0]^2\|}_{K_2} \\ &+ \underbrace{\|\mathbb{E}_N p_i p_i'[p_i'\beta_0 - p_i'\widehat{\beta}]^2\|}_{K_3}\end{aligned}$$

Step 1. K_1. Define an event $\mathcal{E}_N := \{\widehat{\eta}_k \in T_N \quad \forall k \in [K]\}$, such that the nuisance parameter estimate $\widehat{\eta}_k$ belongs to the realization set T_N for each fold $k \in [K]$. By union bound, this event holds w.p $1-o(1)$: $\mathrm{P}_{P_N}(\mathcal{E}_N) \geqslant 1 - K\epsilon_N = 1 - o(1)$.

$$\begin{aligned}K_1 = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{n,k} p_i p_i'[\widehat{Y}_i - Y_i]^2 := \frac{1}{K}\sum_{k=1}^K \big[&\underbrace{\mathbb{E}_{n,k} p_i p_i'[\widehat{Y}_i - Y_i]^2 - \mathbb{E}[p_i p_i'[\widehat{Y}_i - Y_i]^2]|(W_i)_{i\in J_k^c}}_{K_{1,k,1}} \\ &+ \underbrace{\mathbb{E}[p_i p_i'[\widehat{Y}_i - Y_i]^2|(W_i)_{i\in J_k^c}]}_{K_{1,2,k}}\big]\end{aligned}$$

Conditionally on $(W_i)_{i\in J_k}$, the estimator $\widehat{\eta}_k$ is non-stochastic. On the event $\mathcal{E}_N$ let us apply Lemma 6.6 for $w_i := p_i[Y_i(\widehat{\eta}) - Y_i]$. By Assumption 2.6, for $p = q > 2$, $\sup_{\eta\in T_N}\mathbb{E}\|w_i\|^m \leqslant \xi_d(\mathbb{E}|Y_i(\eta) - Y_i(\eta_0)|^q)^{1/q} \leqslant \xi_d$. For d_N defined in Lemma 6.6, conditionally on $(W_i)_{i\in J_k^c}$, $K_{1,k,1} = \|\mathbb{E}_{n,k} w_i w_i' - \mathbb{E}[w_i w_i|(W_i)_{i\in J_k^c}]\| = O_P(d_n) = o_P(1)$. Lemma 6.1 implies that $K_{1,k,1} = O_P(d_n) = o_P(1)$ unconditionally. The sum of K $O_P(d_N)$ terms yields: $\frac{1}{K}\sum_{k=1}^K \|K_{1,k,1}\| = O_P(d_N)$.

$$\begin{aligned}\|K_{1,2,k}\| = \|\mathbb{E}[p_i p_i'[\widehat{Y}_i - Y_i]^2]|(W_i)_{i\in J_k^c}\| &\leqslant \sup_{\alpha:\|\alpha\|=1}\|\mathbb{E}[(\alpha' p_i)p_i'[\widehat{Y}_i - Y_i]|(W_i)_{i\in J_k^c}\|^2 \\ &\leqslant \xi_d^2\|\mathbb{E}[p_i[\widehat{Y}_i - Y_i]|(W_i)_{i\in J_k^c}]\|^2 \\ &\leqslant \xi_d^2/NB_N^2 = o(1) \qquad \text{(Lemma 6.3, Lemma 6.1)}\end{aligned}$$

Step 2. K_2. Set $w_i := p_i[r_i + U_i] = p_i[Y_i - p_i'\beta_0]$. By Assumption 2.6, for $p = m > 2$, $\mathbb{E}\|w_i\|^m \leqslant \xi_d l_d r_d + \xi_d(\mathbb{E}|U_i|^m)^{1/m} \leqslant \xi_d$. For d_N defined in Lemma 6.6, $K_2 = \|\mathbb{E}_N w_i w_i' - \mathbb{E} w_i w_i\| = O_P(d_N) = o_P(1)$. To bound K_3, recognize that

$$\begin{aligned}K_3 &\lesssim_P \|\mathbb{E}_N p_i p_i'\| \max_{1\leqslant i\leqslant N}\|p_i'(\widehat{\beta} - \beta_0)\|^2 \\ &\lesssim_P \|\mathbb{E}_N p_i p_i'\| \max_{1\leqslant i\leqslant N}(\widehat{\beta} - \beta_0)' p_i p_i'(\widehat{\beta} - \beta_0) \\ &\lesssim_P \|\mathbb{E}_N p_i p_i'\|\|\widehat{\beta} - \beta_0\|^2 \max_{1\leqslant i\leqslant n}\|p_i p_i'\| \\ &\lesssim_P [\lambda_{\max} + O_P(\sqrt{\frac{\xi_d^2\log N}{N}})]O_P(\frac{d}{N}\max_{1\leqslant i\leqslant n}\|p_i p_i'\|) = o_P(1) \qquad \text{(Assumption 2.6)}\end{aligned}$$

The bounds on $K_1 - K_3$ conclude the proof of Theorem 2.1[d].

■

Proof of Theorem 2.2(a). Similar to the Equation 6.5, define

$$I_1(x) = \alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{G}_N [p_i(u_i + r_i)]$$

$$I_2(x) = \sqrt{N} \alpha(x)^\top Q^{-1} \mathbb{E}_N p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)]$$

$$I_3(x) = \sqrt{N} \alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{E}_N p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)]$$

Decompose

$$\sqrt{N} \alpha(x)^\top (\widehat{\beta} - \beta_0) = \sqrt{N} \alpha(x)^\top Q^{-1} \mathbb{G}_N [p_i [r_i + u_i]] + I_1(x) + I_2(x) + I_3(x)$$

Step 1. Bound on $I_1(x)$ is shown in Step 1 of Lemma 4.2 of Belloni et al. (2015). We copy their proof for completeness. Let us show that

$$\sup_{x \in \mathcal{X}} |\alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{G}_N [p_i u_i| \lesssim_P N^{1/m} \sqrt{\frac{\xi_d^2 \log^2 N}{N}}$$

Conditional on the data, let $T = \{t_1, t_2, .., t_N\} \in \mathcal{R}^N : t_i = \alpha(x)^\top (\widehat{Q}^{-1} - Q^{-1}) p_i u_i, x \in \mathcal{X}$. Define the norm $\| \cdot \|_{N,2}^2 = \sum_{i=1}^N t_i^2$ on $\mathcal{R}^N$. Let $(\gamma_i)_{i=1}^N$ be independent Rademacher random variables $\mathrm{P}(\gamma_i = 1) = \mathrm{P}(\gamma_i = -1) = \frac{1}{2}$. Symmetrization inequality (i) and Dudley (1967) (ii) imply:

$$\begin{aligned} \mathbb{E} \sup_{x \in \mathcal{X}} |\alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{G}_N p_i u_i| &\leqslant 2 \mathbb{E}_\gamma \sup_{x \in \mathcal{X}} |\alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{G}_N [p_i u_i] | \gamma_i \\ &\leqslant C \int_0^\theta (\log N(\epsilon, T, \| \|_{N,2}))^{1/2} d\epsilon := J \end{aligned}$$

where $N(\epsilon, T, \| \|_{N,2})$ is the covering number of set T and

$$\theta = 2 \sup_{t \in T} \|t\|_{N,2} \leqslant 2 \|\widehat{Q}^{-1} - Q\| \|\widehat{Q}\|^{1/2} \max_{1 \leqslant i \leqslant N} |u_i|$$

Since for any $x \neq x'$:

$$\left[[\alpha(x) - \alpha(x')]^\top [\widehat{Q}^{-1} - Q^{-1}] p_i u_i \right]_{L^2_{P_N}} \leqslant \xi_d^L \|x - x'\| \|\widehat{Q}^{-1} - Q^{-1}\| \|\widehat{Q}\|^{1/2} \max_{1 \leqslant i \leqslant N} |u_i|$$

we have for some $C > 0$:

$$N(\epsilon, T, \| \|_{N,2}) \leqslant \left(\frac{\|\widehat{Q}^{-1} - Q^{-1}\| \|\widehat{Q}\|^{1/2} \max_{1 \leqslant i \leqslant N} |u_i|}{\epsilon} \right)^d$$

$$\int_0^\theta (\log N(\epsilon, T, \|\|_{N,2}))^{1/2} d\epsilon \leqslant \|\widehat{Q}^{-1} - Q\| \|\widehat{Q}\|^{1/2} \max_{1\leqslant i\leqslant N} |u_i| \int_0^2 \sqrt{d} \log^{1/2}(C\xi_d^L/\epsilon) d\epsilon$$

By Assumption 2.6, we have

$$\mathbb{E} \max_{1\leqslant i\leqslant N} |u_i| \leqslant (\mathbb{E}(\max_{1\leqslant i\leqslant N} |u_i|)^m)^{1/m} \leqslant (\mathbb{E}N|u_i|^m)^{1/m} \leqslant N^{1/m}$$

By Assumption 2.2, $\|\widehat{Q} - Q\| \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}}$. Finally, from Assumption 2.8,

$$\log \xi_d^L \lesssim \log d \lesssim \log N$$

$$J \lesssim_P N^{1/m} \sqrt{\frac{\xi_d^2 \log^2 N}{N}}$$

Step 2. Observe that:

$$\sup_{x\in\mathcal{X}} |\alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{G}_N[p_i r_g(x_i)] \lesssim_P \sqrt{\frac{\xi_d^2 \log N}{N}} l_d r_d \sqrt{d}$$

Steps 1 and 2 give the bound on $I_1(x)$. Step 3. Define an event $\mathcal{E}_N := \{\widehat{\eta}_k \in T_N \quad \forall k \in [K]\}$, such that the nuisance parameter estimate $\widehat{\eta}_k$ belongs to the realization set T_N for each fold $k \in [K]$. By union bound, this event holds w.h.p.

$$\mathrm{P}_{P_N}(\mathcal{E}_N) \geqslant 1 - K\epsilon_N = 1 - o(1).$$

Decompose $I_2(x)$ as follows:

$$I_2(x) = \alpha(x)^\top Q^{-1} \mathbb{E}_N p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)] = \frac{1}{K} \sum_{k=1}^K \underbrace{\mathbb{E}_{n,k} \alpha(x)^\top Q^{-1} p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)]}_{I_{2,k}(x)}$$

$$I_3(x) = \alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] \mathbb{E}_N p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)] = \frac{1}{K} \sum_{k=1}^K \underbrace{\mathbb{E}_{n,k} \alpha(x)^\top [\widehat{Q}^{-1} - Q^{-1}] p_i [Y_i(\widehat{\eta}) - Y_i(\eta_0)]}_{I_{3,k}(x)}$$

It suffices to show that $\sup_{x\in\mathcal{X}} |I_{i,k}(x)| = o_P(1) \quad \forall i \in \{2,3\}, k \in [K]$.

Define $\mathbb{G}_{n,k} f(x_i) := \frac{1}{n} \sum_{i\in J_k} f(x_i) - \mathbb{E}\big[f(x_i)|J_k^c\big]$ Define $v_i := p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]$

$$|I_{2,k}(x)| \leqslant \sup_{x\in\mathcal{X}} \sqrt{N} \underbrace{|\alpha(x)^\top Q^{-1} \mathbb{G}_{n,k} v_i|}_{I_{21,k}(x)} + \sup_{x\in\mathcal{X}} \underbrace{\sqrt{N} |\alpha(x)^\top Q^{-1} \mathbb{E} v_i| (W_i)_{i\in J_k}|}_{I_{22,k}(x)}$$

$$|I_{22,k}(x)| \leqslant \|\alpha\| \|Q^{-1}\| \sup_{\eta\in T_N} \|\sqrt{N} \mathbb{E} v_i| (W_i)_{i\in J_k}\| \lesssim \sup_{\eta\in T_N} \|\sqrt{N} \mathbb{E} v_i\| = B_N = o(1)$$

(Assumption 2.5)

To bound $\sup_{x\in\mathcal{X}} I_{21,k}(x)$, we apply Lemma 6.2 conditionally on J_k^c so that $\widehat{\eta}_k$ can be treated as fixed. Consider a function class

$$\mathcal{F} := \{f_i = \alpha(x)Q^{-1}v_i, x \in \mathcal{X}\}$$

with an square integrable envelope $F := \|\alpha(x)\|\|v_i\| = \|v_i\|$:

$$\|F\|_{L_P^2} = \mathbb{E}\|v_i\|^2 \leqslant \|Q^{-1}\|^2\xi_d^2 \sup_{\eta\in T_N} \mathbb{E}(Y_i(\eta) - Y_i(\eta_0))^2 \leqslant \|Q^{-1}\|^2\xi_d^2\kappa_N^2$$

Define the second moment bound $\sigma^2 := \mathbb{E}\|v_i\|^2 = O(\xi_d^2\kappa_N^2)$ which uniformly bounds the second moment of every element in $\mathcal{F}$

$$\sup_{f\in\mathcal{F}} \mathbb{E}f^2 \leqslant \sigma^2 = \|v_i\|_{L_P^2}^2$$

To determine the bracket size, recognize that

$$|[\alpha(x) - \alpha(x')]^\top Q^{-1}v_i| \leqslant \xi_d^L\|x - x'\|\|Q^{-1}\|\|v_i\|$$

and therefore

$$\sup_Q N(\mathcal{F}, L^2(Q), \epsilon\|F\|_{L_Q^2}) \leqslant (\frac{\xi_d^L/\lambda_{\min}}{\epsilon})^r$$

Plugging in $\sigma = \|F\|_{L_P^2} = \xi_d\kappa_N$, $A = \xi_d^L/\lambda_{\min}$, $V = r$ into Lemma 6.2, we obtain:

$$\sup_{x\in\mathcal{X}} |\alpha(x)^\top Q^{-1}\mathbb{G}_{n,k}p_i[Y_i(\widehat{\eta}) - Y_i(\eta_0)]| \lesssim_P \xi_d\kappa_N\sqrt{\log \xi_d^L} + N^{-1/2+1/q}\log \xi_d^L \tag{6.7}$$

$$\lesssim_P \xi_d\kappa_N\sqrt{\log N} + N^{-1/2+1/q}\log N \tag{6.8}$$

Lemma 6.1 implies that the bounds on $I_{21,k}(x)$ and $I_{22,k}(x)$ are unconditional. Therefore,

$$\begin{aligned}\sup_{x\in\mathcal{X}} |I_{21}(x)| &\lesssim_P \frac{1}{K}\sum_{k=1}^K[\sup_{x\in\mathcal{X}} |I_{21,k}(x)| + \sup_{x\in\mathcal{X}} |I_{22,k}(x)|] \\ &\lesssim_P \xi_d\kappa_N\sqrt{\log N} + N^{-1/2+1/q}\log N + B_N\end{aligned}$$

Step 4. We wish to bound $I_3(x)$.

$$\begin{aligned}|I_{3,k}(x)| &\leqslant \sup_{x\in\mathcal{X}} |\alpha(x)^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{G}_{n,k}v_i| + \sup_{x\in\mathcal{X}} \sqrt{N}|\alpha(x)^\top[\widehat{Q}^{-1} - Q^{-1}]\mathbb{E}[v_i|J_k^c]| \\ &\leqslant \|\widehat{Q}^{-1} - Q^{-1}\|\|\mathbb{G}_{n,k}v_i\| + \sup_{x\in\mathcal{X}} \|\alpha(x)\|\sqrt{\frac{\xi_d^2\log N}{N}}B_N \\ &\leqslant \sqrt{\frac{\xi_d^2\log N}{N}}(\xi_d\kappa_N\sqrt{\log N} + N^{-1/2+1/q}\log N) + O_P(\sqrt{\frac{\xi_d^2\log N}{N}}B_N)\end{aligned}$$

Theorem 2.2(b) is established in Belloni et al. (2015).Proof of Corollary 2.1 follows from

Theorem 2.2 (a) and Theorem 4.4 of Belloni et al. (2015).

■

Proof of Corollary 3.1. Let us show that Y_η given by Equation 1.3 is an efficient signal that satisfies Assumption 2.5

$$
\begin{aligned}
Y_\eta - Y_{\eta_0} = & \underbrace{[\mu(1,Z)-\mu_0(1,Z)][1-\frac{D}{s_0(Z)}]}_{S_1} - \underbrace{[\mu(0,Z)-\mu_0(0,Z)][1-\frac{1-D}{1-s_0(Z)}]}_{S_1'} \\
& + \underbrace{[s_0(Z)-s(Z)]\big[[Y^1-\mu_0(1,Z)]\frac{D}{s(Z)s_0(Z)}}_{S_2} - \underbrace{[Y^0-\mu_0(0,Z)]\frac{1-D}{(1-s(Z))(1-s_0(Z))}]}_{S_2'} \\
& + \underbrace{[s_0(Z)-s(Z)]\big[[\mu(1,Z)-\mu_0(1,Z)]\frac{D}{s(Z)s_0(Z)}}_{S_3} - \underbrace{[\mu(1,Z)-\mu_0(0,Z)]\frac{1-D}{(1-s(Z))(1-s_0(Z))}]}_{S_3'}
\end{aligned}
$$

Let us see that Assumption 2.5 (1) is satisfied with $B_N := \xi_d \mathbf{m}_N \mathbf{s}_N$. The terms $S_i, S_i', i \in \{1,2\}$ are mean zero conditionally on Z. Assumptions 1.1 and 1.2 imply that any technical regressor $p(X)$ is uncorrelated with $S_1 + S_1' + S_2 + S_2'$.

$$
\begin{aligned}
\mathbb{E}\big[[S_1+S_1'+S_2+S_2']|Z\big] = 0 \quad \Rightarrow \quad & \mathbb{E}p(X)\big[[S_1+S_1'+S_2+S_2']\big] = 0 \\
\|\mathbb{E}p(X)[S_3+S_3']\|^2 = & \sum_{j=1}^{d}(\mathbb{E}p_j(X)[S_3+S_3'])^2 \leqslant \sum_{j=1}^{d}\mathbb{E}p_j^2(X)\mathbb{E}(S_3+S_3')^2 \\
& \leqslant \xi_d^2(\mathbf{m}_N)^2(\mathbf{s}_N)^2
\end{aligned}
$$

Assumption 2.5 (2) is satisfied with $\kappa_N := [\mathbf{m}_N^1 \vee \mathbf{m}_N \vee \mathbf{s}_N]$.

$$
\begin{aligned}
\mathbb{E}S_1^2 &= \mathbb{E}[\mu(1,Z)-\mu_0(1,Z)]^2\mathbb{E}\big[[1-\frac{D}{s_0(Z)}]^2|Z=z\big] \lesssim (\mathbf{m}_N)^2 \\
\mathbb{E}(S_1')^2 &= \mathbb{E}[\mu(0,Z)-\mu_0(0,Z)]^2\mathbb{E}\big[[1-\frac{1-D}{1-s_0(Z)}]^2|Z=z\big] \lesssim (\mathbf{m}_N)^2 \\
\mathbb{E}S_2^2 + (S_2')^2 &= \mathbb{E}[s(Z)-s_0(Z)]^2\mathbb{E}\Big[\big[[Y^1-\mu_0(1,Z)]^2|Z=z\big]\frac{1}{s_0(Z)s^2(Z)} \\
&\quad + [Y^0-\mu_0(0,Z)]^2|Z=z\big]\frac{1}{1-s_0(Z)(1-s(Z))^2}\Big] \lesssim \mathbf{s}_N^2\frac{1}{\bar{s}_0^3} \\
\mathbb{E}S_3^2 + (S_3')^2 &\lesssim \min(\mathbf{m}_N^2, \mathbf{s}_N^2)\frac{1}{\bar{s}_0^3}
\end{aligned}
$$

To sum up,

$$
\xi_d(\mathbb{E}(Y_\eta - Y_{\eta_0})^2)^{1/2} \lesssim_P \xi_d(\mathbf{m}_N \vee \mathbf{s}_N)
$$

■

Proof of Corollary 3.2. Let us show that Y_η given by Equation 1.4 is an efficient signal that

satisfies Assumption 2.5

$$Y_\eta - Y_{\eta_0} = \underbrace{[\mu(Z) - \mu_0(Z)][1 - \frac{D}{s_0(Z)}]}_{S_1} + \underbrace{[s_0(Z) - s(Z)][Y - \mu_0(Z)]\frac{D}{s(Z)s_0(Z)}}_{S_2}$$
$$+ \underbrace{[s_0(Z) - s(Z)][\mu(Z) - \mu_0(Z)]\frac{D}{s(Z)s_0(Z)}}_{S_3}$$

Let us see that Assumption 2.5 (1) is satisfied with $B_N := \xi_d \mathbf{m}_N \mathbf{s}_N$. The terms S_1, S_2 are mean zero conditionally on Z. Assumptions 1.3 and 1.2 imply that any technical regressor $p(X)$ is uncorrelated with $S_1 + S_2$.

$$\|\mathbb{E}p(X)S_3\| \lesssim \xi_d \mathbf{m}_N \mathbf{s}_N$$

$$\mathbb{E}S_1^2 = \mathbb{E}[\mu(Z) - \mu_0(Z)]^2 \mathbb{E}\big[[1 - \frac{D}{s_0(Z)}]^2 | Z = z\big] \lesssim \mathbf{m}_N^2$$
$$\mathbb{E}S_2^2 = \mathbb{E}[s(Z) - s_0(Z)]^2 \mathbb{E}\big[[Y - \mu_0(Z)]^2 | Z = z\big] \frac{1}{s_0(Z)s^2(Z)} \lesssim \mathbf{s}_N^2 \frac{1}{\bar{s}_0^3}$$
$$\mathbb{E}S_3^2 = \mathbb{E}[s_0(Z) - s(Z)]^2 [\mu(Z) - \mu_0(Z)]^2 \frac{1}{s^2(Z)s_0(Z)} \lesssim \min(\mathbf{m}_N^2, \mathbf{s}_N^2) \frac{1}{\bar{s}_0^3}$$

To sum up,

$$\xi_d (\mathbb{E}(Y_\eta - Y_{\eta_0})^2)^{1/2} \lesssim_P \xi_d [\mathbf{m}_N \vee \mathbf{s}_N]$$

■

Proof of Corollary 3.3. Let us show that Y_η given by Equation 1.4 is an efficient signal that satisfies Assumption 2.5

$$Y_\eta - Y_{\eta_0} = \underbrace{-\partial_w \log f_0(W|X)[\mu(X,W) - \mu_0(X,W)] + \partial_w[\mu(X,W) - \mu_0(X,W)]}_{S_1}$$
$$+ \underbrace{[\partial_w \log f_0(W|X) - \partial_w \log f(W|X)][Y - \mu_0(X,W)]}_{S_2}$$
$$+ \underbrace{[\partial_w \log f_0(W|X) - \partial_w \log f(W|X)][\mu(X,W) - \mu_0(X,W)]}_{S_3}$$

Let us see that Assumption 2.5 (1) is satisfied with $B_N := \xi_d \mathbf{f}_N \mathbf{m}_N$. The terms S_1, S_2 are mean zero conditionally on Z. Since $X \subset Z$, any technical regressor $p(X)$ is uncorrelated with $S_1 + S_2$

$$\|\mathbb{E}p(X)S_3\| = \|\mathbb{E}p(X)S_3\| \lesssim \xi_d \mathbf{f}_N \mathbf{m}_N$$

$$\mathbb{E}S_1^2 \leqslant \mathbb{E}(-\partial_w \log f_0(W|X)[\mu(X,W)-\mu_0(X,W)])^2 + \mathbb{E}(\partial_w[\mu(X,W)-\mu_0(X,W)])^2 \lesssim \mathbf{m}_N^2$$
$$\mathbb{E}S_2^2 \leqslant (\mathbb{E}\big[[Y-\mu_0(X,W)]^2|X,W\big])\mathbf{f}_N^2 \lesssim \mathbf{f}_N^2$$
$$\mathbb{E}S_3^2 \leqslant (\mathbb{E}[\mu(X,W)-\mu_0(X,W)]^2)\mathbf{f}_N^2 \lesssim \min[\mathbf{f}_N^2 \vee \mathbf{m}_N^2]$$

$$\xi_d(\mathbb{E}(Y_\eta - Y_{\eta_0})^2)^{1/2} \lesssim_P \xi_d[\mathbf{f}_N \vee \mathbf{m}_N]$$

■

6.4 Proof of Example 5

We establish the argument in the following steps.

Proof. Define $\tilde{Y} = DY$, $\tilde{p}_\mu(Z) = Dp_Z(Z)$, $\tilde{r}_\mu(Z) = Dr_\mu(Z)$, and $\tilde{\epsilon} = D[Y-\mu(Z)]$. Step 1. Let us show that the original coefficient θ, defined in Equation (3.1), satisfies

$$\tilde{Y} = \tilde{D}'\theta + \tilde{r}_\mu(Z) + \tilde{\epsilon}, \quad \mathbb{E}[\tilde{\epsilon}|\tilde{D}] = 0$$

Indeed,

$$\begin{aligned}\mathbb{E}[\tilde{Y}|D,Z] &= \mathbb{E}[DY|D,Z] = D\mathbb{E}[Y|Z,D] = D[\mu_0(Z)] = Dp_\mu(Z)'\theta + Dr(Z)\\ &= \tilde{p}_\mu(Z)'\theta + \tilde{r}_\mu(Z)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\tilde{\epsilon}|\tilde{D}] &= \mathbb{E}\big[[\tilde{Y} - \tilde{p}_\mu(Z)'\theta - \tilde{r}_\mu(Z)]|\tilde{D}\big]\\ &= \mathbb{E}[\mathbb{E}\big[[\tilde{Y} - \tilde{p}_\mu(Z)'\theta - \tilde{r}_\mu(Z)]|D,Z\big]|\tilde{D}]\\ &= 0\end{aligned}$$

Step 2. Recognize that Assumption 3.3 implies that an analog of Assumption 3.3 holds for $\tilde{p}_\mu(Z), \tilde{Y}$. Assumption 3.3 (a) directly assumes bounded Restricted Sparse Eigenavalues for Observed Regressors $\tilde{p}_\mu(Z) = p_\mu(Z)D$. Assumption 3.3 (b) is satisfied with $\mathbb{E}\tilde{p}_{\mu,j}(Z)^2 = \mathbb{E}Dp_{\mu,j}^2(Z) \geqslant \underline{s}\mathbb{E}p_{\mu,j}^2(Z) =: c'$ where $c' := c\underline{s}$ is the new lower bound on the moments of $\mathbb{E}\tilde{p}_{\mu,j}(Z)$ for observed regressors. Assumption 3.3 (c) is satified with the $\mathbb{E}\tilde{r}_\mu(Z)^2 \leqslant \mathbb{E}r_\mu(Z)^2 \leqslant Cs\log(p\vee N)/N$.

■

References

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernandez-Val, I. (2011). Conditional quantile processes based on series or many regressors.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2013). Program evaluation and causal inference with high-dimensional data. *arXiv preprint arXiv:1311.2645*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Journal of Economic Perspectives*, 28(2).

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Blundell, R., Horowitz, J., and Parey, M. (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, (3):29–51.

Bühlmann, P. and van der Geer, S. (2011). Statistics for high-dimensional data. *Springer Series in Statistics*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., et al. (2016a). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. (2016b). Locally robust semiparametric estimation.

Graham, B. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2):437–452.

Graham, B., Pinto, C., and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3):1053 – 1079.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Hansen, B. (2014). A unified asymptotic distribution theory for parametric and non-parametric least squares.

Hausman, J. and Newey, W. (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, (63):1445–1476.

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

Horwitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47(260).

Luo, Y. and Spindler, M. (2016). High-dimensional l2 boosting: Rate of convergence. *arXiv:1602.08927*.

Newey, W. (2007). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Newey, W. and Stoker, T. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.

Portnoy, S. and Koenker, R. (1989). Adaptive l-estimation for linear models. *The Annals of Statistics*.

Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association*, 90(429):122–129.

Robins, J., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 89(427):846–866.

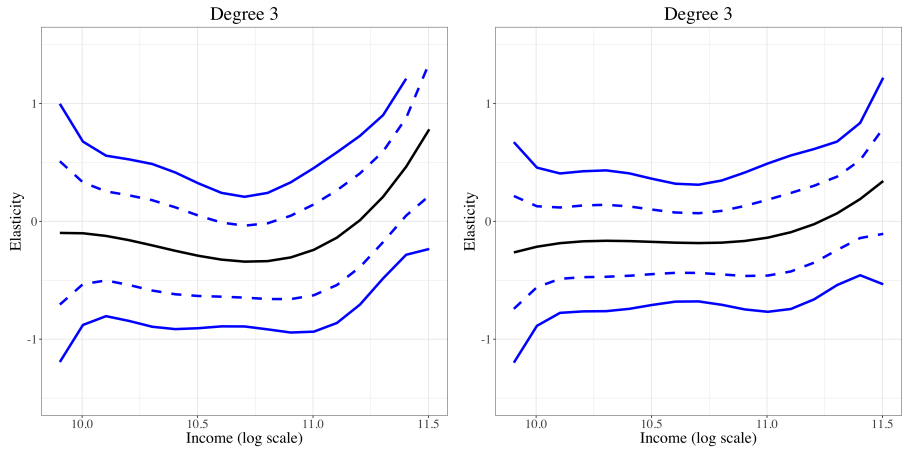
Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72.

Schmalensee, R. and Stoker, T. (1999). Household gasoline demand in the united states. *Econometrica*, (67):645–662.

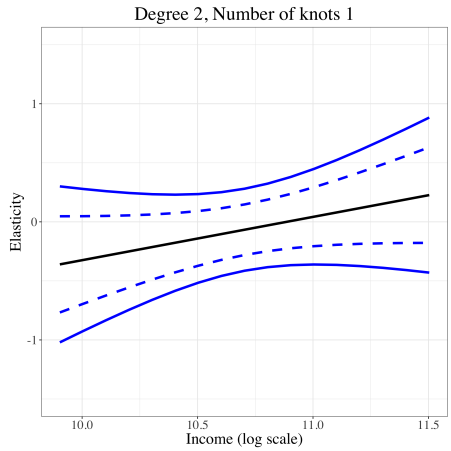
Wager, S. and Athey, S. (2016). Estimation and inference of heterogeneous treatment effects using random forests. *https://arxiv.org/abs/1510.04342*.

Yatchew, A. and No, J. A. (2001). Household gasoline demand in canada. *Econometrica*, pages 1697–1709.

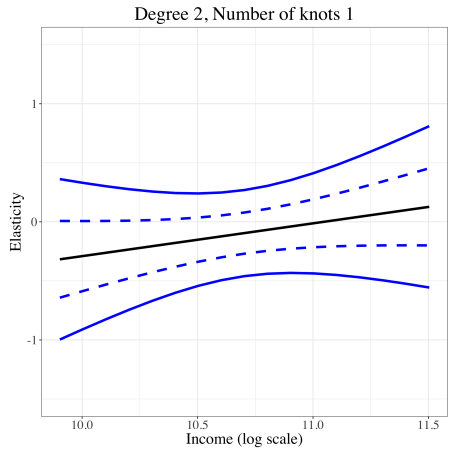


(a) Step 2: $l(x, z)$ is estimated by Lasso. Step 3: polynomials of degree 3.

(b) Step 2: $l(x, z)$ is estimated by random forest. Step 3: polynomials of degree 3.

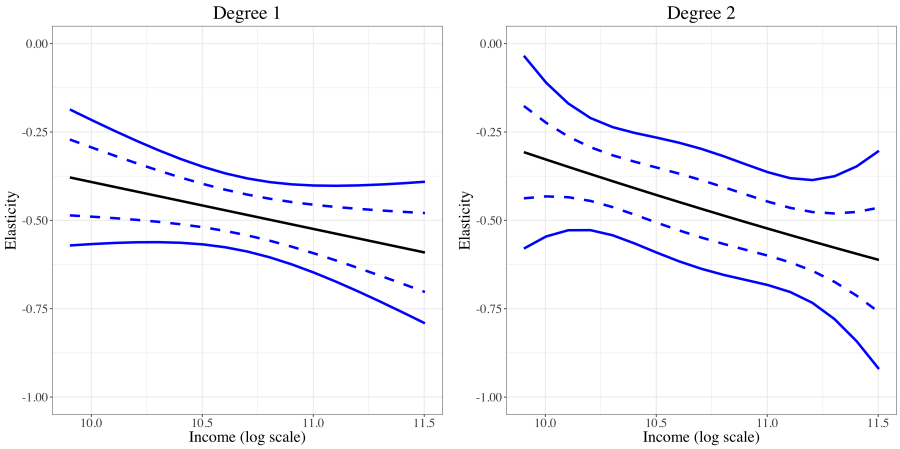


(c) Step 2: $l(x, z)$ is estimated by Lasso. Step 3: B-splines of order 2 with 1 knot.



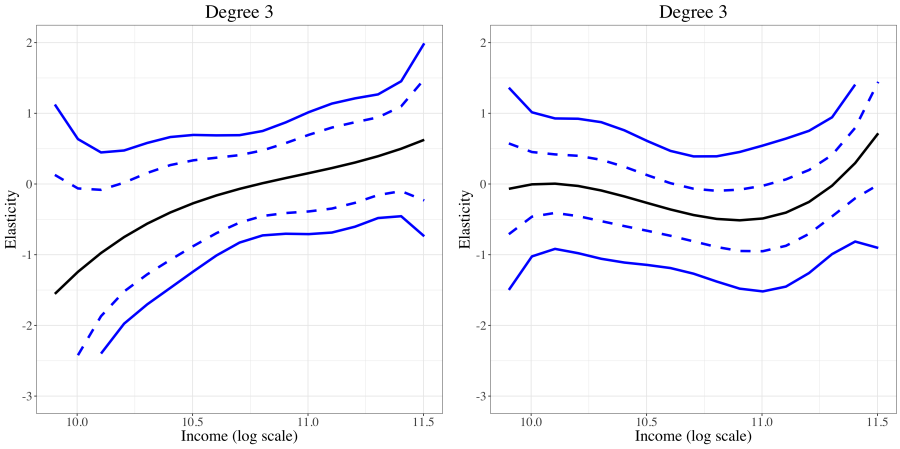
(d) Step 2: $l(x, z)$ is estimated by random forest. B-splines of order 2 with 1 knot.

Figure 1: 95% confidence bands for the best linear approximation of the average price elasticity conditional on income with accounting for the demographic controls in the first stage. The black line is the estimated function, the dashed(solid) blue lines are the pointwise (uniform) confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(w, x, z)$, (2) second-stage estimation of the conditional density $f(w|x, z)$, and (3) third-stage estimation of the target function $g(x)$ by least squares series. Step 1 is performed using Lasso with standardized covariates and the penalty choice $\lambda = 2.2\sqrt{n}\widehat{\sigma}\Phi^{-1}(1-\gamma/2p)$, where $\gamma = 0.1/\log n$ and $\widehat{\sigma}$ is the estimate of the residual variance. Step 2 is performed by estimating the regression function of $l(x, z) = \mathbb{E}[W|X = x, Z = z]$ and estimating the density $f(w - l(x, z))$ of the residual $w - l(x, z)$ by adaptive kernel density estimator of Portnoy and Koenker (1989) with the Silverman choice of bandwidth. The regression function $l(x, z)$ is estimated lasso (1a, 1c) and random forest (1b, 1d). Step 3 is performed using B-splines of order 2 with the number of knots equal to one (1c, 1d) and polynomial functions of order 3. (1a, 1b). $B = 200$ weighted bootstrap repetitions.



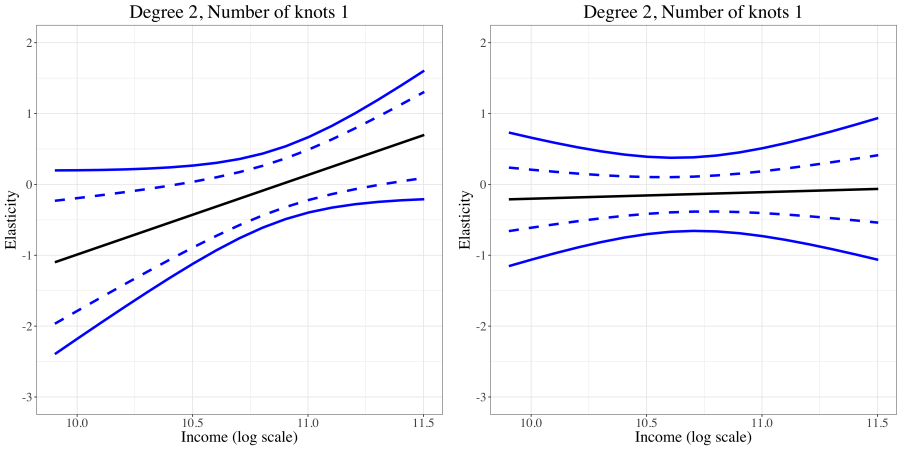
(a) Polynomial degree $q = 1$ (b) Polynomial degree $q = 2$

Figure 2: 95% confidence bands for the best linear approximation of the average price elasticity conditional on income without accounting for the demographic controls in the first stage. The black line is the estimated function, the dashed blue lines and the solid blue lines are the pointwise and the uniform confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(w, x) = \mathbb{E}[Y|W = w, X = x]$, (2) second-stage estimation of the conditional density $f(W = w|X = x)$, and (3) third-stage estimation of the target function $g(x)$ by least squares series. Step 1 is performed using least squares series regression using polynomial functions $\{1, x, \ldots, x^q\}, q = 3$ whose power q is chosen by cross-validation out of $\{1, 2, 3\}$. Step 2 is performed by kernel density estimator with the Silverman choice of bandwidth. Step 3 is performed using polynomial functions $\{1, x, \ldots, x^q\}$ and is shown for $q = 1$ and $q = 2$. $B = 200$ weighted bootstrap repetitions.



(a) Large Households, Polynomials of degree 3.

(b) Small Households, Polynomials of degree 3.



(c) Large Households, B-splines of degree 2 with 1 knot.

(d) Small Households, B-splines of degree 2 with 1 knot.

Figure 3: 95% confidence bands for the best linear approximation of the average price elasticity conditional on income with accounting for the demographic controls in the first stage by household size. The black line is the estimated function, the dashed(solid) blue lines are the pointwise (uniform) confidence bands. The estimation algorithm has three steps: (1) first-stage estimation of the conditional expectation function $\mu(w, x, z)$, (2) second-stage estimation of the conditional density $f(w|x, z)$, and (3) third-stage estimation of the target function $g(x)$ by least squares series. Step 1 is performed using Lasso with standardized covariates and the penalty choice $\lambda = 2.2\sqrt{n}\widehat{\sigma}\Phi^{-1}(1 - \gamma/2p)$, where $\gamma = 0.1/\log n$ and $\widehat{\sigma}$ is the estimate of the residual variance. Step 2 is performed by estimating the regression function of $l(x, z) = \mathbb{E}[W|X = x, Z = z]$ and estimating the density $f(w - l(x, z))$ of the residual $w - l(x, z)$ by adaptive kernel density estimator of Portnoy and Koenker (1989) with the Silverman choice of bandwidth. The regression function $l(x, z)$ is estimated lasso. Step 3 is performed using B-splines of order 2 with the number of knots equal to one (3c, 3d) and using non-orthogonal polynomial functions of degree 3 (3a, 3b). $B = 200$ weighted bootstrap repetitions.