MACHINE LEARNING UNDER ENDOGENEITY

Edvard Bakhitov

A DISSERTATION

in

Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Amit Gandhi, Vice President and Technical Fellow, Airbnb

Graduate Group Chairperson

David Dillenberger, Professor of Economics

Dissertation Committee

Xu Cheng, Associate Professor of Economics

Francis X. Diebold, Professor of Economics

Karun Adusumilli, Assistant Professor of Economics

MACHINE LEARNING UNDER ENDOGENEITY

ACKNOWLEDGEMENT

I am especially indebted to my advisor, Amit Gandhi, for his vast knowledge, invaluable guidance, and continuous support throughout my time at Penn. I would like to express my deep gratitude to my committee members, Karun Adusumilli, Xu Cheng and Frank Diebold, whose invaluable input greatly helped to shape my research agenda. I also wish to extend my special thanks to Wayne Gao for his insightful comments and suggestions.

I am thankful to my colleagues and friends I have met at Penn. Among them, Minji Bang, Görkem Bostanci, Philippe Goulet Coulombe, Paul Décaire, Max Esser, Changhwa Lee, Hanbaek Lee, Boyuan Zhang, and Ewelina Žurowska. I deeply thank Amandeep Singh for playing a big part in our joint projects. I am also grateful to Jing Tao for her guidance and support during my research assistant years.

I wish to express my great appreciation to Christos Koulovatianos and Gautam Tripathi who helped me open a door to Penn.

Finally, I thank my family and friends for their everlasting support ever since I started my PhD.

ABSTRACT

MACHINE LEARNING UNDER ENDOGENEITY

Edvard Bakhitov

Amit Gandhi

Recent advances in machine learning literature provide a series of new algorithms that both address endogeneity and can be applied in high-dimensional environments, we call them MLIV. These algorithms are data-driven and exploit various forms of regularization to ameliorate the ill-posedness of the problem while maintaining the functional form flexibility. In this thesis, we discuss how MLIV estimators can be used to answer economic questions.

In the first chapter, *Causal Gradient Boosting: Boosted Instrumental Variables Regression*, we propose an MLIV algorithm called boostIV that builds on the traditional gradient boosting algorithm and corrects for the endogeneity bias. The algorithm is very intuitive and resembles an iterative version of the standard 2SLS estimator. The second chapter, *Automatic Debiased Machine Learning in Presence of Endogeneity*, introduces an approach for performing valid asymptotic inference on regular functionals of MLIV estimators. The approach is based on construction of an orthogonal moment function that has a zero derivative with respect to the MLIV estimator. We develop a penalized GMM estimator of the bias correction term necessary to obtain asymptotically normal debiased estimates and derive its convergence rate. We also give conditions for root-n consistency and asymptotic normality of the debiased MLIV estimator of the functional of interest. Finally, in the third chapter, *Flexible Demand Estimation using Machine Learning*, we demonstrate how to estimate substitution patterns in the market for sodas using the debiasing procedure from the second chapter.

These three chapters are highly interconnected. The first chapter proposes a new MLIV algorithm for flexible estimation in presence of endogenous regressors. However, it focuses on the underlying structural function which in the majority of cases does not have a clear economic interpretation. While the second chapter develops a method to perform inference on functionals of MLIV estimators, which have a clear economic interpretation and can be used to answer various economic questions of interest. Finally, the third chapter investigates an important applied question of flexible estimation of demand for differentiated goods, which is a perfect example of a high-dimensional problem with endogenous regressors. As a result, we get a full picture about the potential of MLIV methods in economics.

Contents

List of Tables

List of Figures

CHAPTER 1: Causal Gradient Boosting: Boosted Instrumental Variables Regression

"Human intelligence discovered a way of perpetuating itself, one not only more durable and more resistant than architecture, but also simpler and easier. Architecture was dethroned. The stone letters of Orpheus gave way to the lead letters of Gutenburg. The book will kill the edifice."

- Victor Hugo

1.1. Introduction

Gradient boosting method is considered one of the leading machine learning (ML) algorithms for supervised learning with structured data. There is a large body of evidence showing that gradient boosting dominates in a significant number of ML competitions conducted on Kaggle[1]. However, recent literature (e.g., see Hartford et al., 2017*a*) has shown that traditional supervised machine learning methods do not perform well in the presence of endogeneity in the explanatory variables.

A common approach to correct for the endogeneity bias is to use instrumental variables (IVs). Nonparametric instrumental variables (NPIV) techniques have gained popularity among applied researchers over the last decade as they do not require imposing (possibly) implausible parametric assumptions on the target function. On the other hand, existing nonparameteric estimation techniques require the researcher to specify a target function approximation (ideally driven by some ex-ante understanding of the data generating process), e.g. a sieve space, which in turn drives the choice of unconditional moment restrictions (or simply put, the choice of IV basis functions). Moreover, the complexity of both modeling and estimation explodes when there are more than a handful of inputs.

In this Chapter, we introduce an algorithm that allows to learn the target function in the presence of endogenous explanatory variables in a data driven way, meaning that the researcher does not have to make a stance on neither the form of the target function approximation nor the choice of instruments. We build on gradient boosting algorithm to transform the standard NPIV problem into a learning problem that accounts for endogeneity in explanatory variable, and thus, we call our algorithm *boostIV*.

We also consider a couple extensions to the boostIV algorithm that might improve its finite sample performance. First, we show how to incorporate optimal IVs, i.e. IVs that achieve the lowest asymptotic variance (Chamberlain, 1987). Second, we augment the boostIV al-

[1]For reference see https://www.kaggle.com/dansbecker/xgboost

gorithm with a post-processing step where we re-estimate the weights on the learnt basis functions, we call this algorithm *post-boostIV*. The idea is based on Friedman and Popescu (2003) who propose to learn an ensemble of basis functions and then apply lasso to perform basis function selection.

To avoid potentially severe finite sample bias due to the double use of data, we resort to the cross-fitting idea of Chernozhukov, Newey and Robins (2018). For the boostIV algorithm we split the data to learn instruments and basis functions on different data folds. We add an additional layer of cross-fitting to the post-boostIV algorithm to update the weights on the learnt basis functions.

Our method has a number of advantages over the standard NPIV approach. First, our approach allows the researcher to be completely agnostic to the choice of basis functions and IVs. Both basis functions and instruments are learnt in a data driven way which picks up the underlying data structure. Second, the method becomes even more attractive when the dimensionality of the problem grows, as the standard NPIV methods suffer greatly from the curse of dimensionality. Intuitively, learning via boosting should be able to construct basis functions that approximately represent the underlying low dimensional data features. However, our approach does not work in purely high-dimensional settings where the number of regressors exceeds the number of observations.

We study the performance of boostIV and post-boostIV algorithms in a series of Monte Carlo experiments. We compare the performance of our algorithms to both the standard sieve NPIV estimator and a variety of modern ML estimators. Our results demonstrate that boostIV performs at worst on par with the state of the art ML estimators. Moreover, we find no empirical evidence that post-boostIV achieves superior performance compared to boostIV and vice versa. However, adding the post-processing step reduces the amount of boosting iterations needed for the algorithm to converge rendering it (potentially) computationally more efficient[2].

This paper brings together two strands of literature. First, our approach contributes to the literature on nonparametric instrumental variables modeling. Newey and Powell (2003) propose to replace the linear relationships in standard linear IV regression with linear projections on a series of basis functions (also see Blundell et al. (2007) for an application to Engel-curve estimation). Darolles et al. (2011) and Hall and Horowitz (2005) suggest to nonparametrically estimate the conditional distribution of endogenous regressors given

[2]To be more precise, there is a trade-off at play. One boostIV iteration takes less time than one post-boostIV iteration as the latter algorithm includes an additional estimation step plus one more layer of cross-fitting. As a result, if adding the post-processing step reduces the amount of boosting iterations significantly, then we achieve computational gains. It might not be the case otherwise.

the instruments, $F(x|z)$, using kernel density estimators. However, despite their simplicity and flexibility, both approaches are subject to the curse of dimensionality. Machine learning literature has recently also contributed to the nonparametric IV literature. Hartford et al. (2017*a*) propose a DeepIV estimator which first estimates $F(x|z)$ with a mixture of deep generative models on which then the structural function is learned with another deep neural network. Kernel IV estimator of Singh et al. (2019) exploits conditional mean embedding of $F(x|z)$, which is then used in the second stage kernel ridge regression. Muandet et al. (2019) avoid the traditional two stage procedure by focusing on the dual problem and fitting just a single kernel ridge regression.

Second, we exploit insights from the boosting literature. Originally boosting came out as an ensemble method for classification in the computational learning theory (Schapire, 1990; Freund, 1995; Freund and Schapire, 1997). Later on Friedman et al. (2000) draw connections between boosting and statistical learning theory by viewing boosting as an approximation to additive modeling. A different perspective on boosting as a gradient descent algorithm in a function space that connects boosting to the more common optimization view of statistical inference (Breiman, 1998, 1999; Friedman, 2001). L_2-boosting introduced by Bühlmann and Yu (2003) provides a powerful tool to learning regression functions. A comprehensive boosting review can be found in Bühlmann and Hothorn (2007).

The remainder of the Chapter is organized as follows. Section 1.2 briefly introduces the NPIV framework. Section 1.3 describes the standard boosting procedure. We present boostIV and post-boostIV in Section 1.4. Section 1.5 talks about hyperparameter tuning. Section 1.6 discusses consistency. We illustrate the numerical performance of our algorithms in Section 1.7. Section 1.8 concludes. All the proofs and mathematical details are left for the Appendix.

1.2. Set-up

Consider the standard conditional mean model of Newey and Powell (2003)

$$y = f(x) + \varepsilon, \quad \mathbb{E}[\varepsilon|z] = 0, \tag{1.1}$$

where y is a scalar random variable, f is an unknown structural function of interest, x is a $d_x \times 1$ vector of (potentially) endogenous explanatory variables, z is a $d_z \times 1$ vector of instrumental variables, and ε is an error term[3]. Suppose that the model is identified and the completeness condition holds, i.e. for all measurable real functions δ with finite

[3]The approach can easily be extended to cases where only some of the regressors are endogenous. Suppose $x = (x_1, x_2)$ where x_1 consists of endogenous regressors and x_2 is a vector of exogenous regressors. Let w be a vector of excluded instruments and set $z = (w, x_2)$. This perfectly fits into the model described by (1.1).

expectation,

$$\mathbb{E}[\delta(x)|z] = 0 \Rightarrow \delta(x) = 0.$$

Intuitively this condition implies that there is enough variation in the instruments to explain the variation in x.

The conditional expectation of (1.1) yields the integral equation

$$\mathbb{E}[y|z] = \int f(x)dF(x|z), \tag{1.2}$$

where F denotes the conditional cdf of x given z. Solving for f directly is an ill-posed problem as it involves inverting linear compact operators (e.g., see Kress, 1989). Note that the model in (1.1) does not have an explicit reduced form, i.e. a functional relationship between endogenous and exogenous variables, however, it is implicitly embedded in F. Thus, from the estimation perspective we have two objects to estimate: (i) the conditional cdf $F(x|z)$ and (ii) the structural function f.

A common approach in applied work is to assume that the relationships between y and x as well as x and z are linear, which leads to a standard 2SLS estimator. However, it can be a very restrictive assumption in practice, which can result in misspecification bias. A lot of more flexible non-parametric extension to 2SLS have been developed in the econometrics literature. The standard approach is ti use the series estimator of Newey and Powell (2003) who propose to replace the linear relationships with a linear projections on a series of basis functions.

To illustrate the approach let us approximate f with a series expansion

$$f(x) \approx \sum_{\ell=1}^{L} \gamma_\ell p_\ell(x),$$

where $p^L(x) = (p_1(x), \ldots, p_L(x))$ is a series of basis functions. It allows us to rewrite the conditional expectation of y given z as

$$\mathbb{E}[y|z] \approx \sum_{\ell=1}^{L} \gamma_\ell \mathbb{E}[p_\ell(x)|z]. \tag{1.3}$$

Let $q^K(z) = (q_1(z), \ldots, q_K(z))$ be a series of IV basis functions. This implies a 2SLS type estimator of γ

$$\hat{\gamma} = \left(\hat{\mathbb{E}}[p^L(x)|z]'\hat{\mathbb{E}}[p^L(x)|z]\right)^{-} \hat{\mathbb{E}}[p^L(x)|z]'y, \tag{1.4}$$

where $\hat{\mathbb{E}}[p^L(x)|z] = q^K(z)\left(q^K(z)'q^K(z)\right)^{-} q^K(z)'p^L(x)$. Given $L,K \to \infty$ as $n \to \infty$, asymptotically one can recover the true structural function. However, in finite samples one has to truncate the sieve at some value. Despite that, the finite sample performance of the estimator hinges crucially on the choice of the approximating space, especially in high dimensions. Moreover, NPIV estimators suffer greatly from the curse of dimensionality which renders them inapplicable in many applications. Alternatively, we propose a data-driven approach, which is agnostic to the choice of the approximating space.

1.3. Revisiting Gradient Boosting

Boosting is a greedy algorithm to learn additive basis function models of the form

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \varphi(x;\theta_m), \tag{1.5}$$

where φ_m are generated by a simple algorithm called a *weak learner* or *base learner*. The weak learner can be any classification or regression algorithm, such as a regression tree, a random forest, a simple single-layer neural network, etc. One could boost the performance (on the training set) of any weak learner arbitrarily high, provided the weak learner could always perform slightly better than chance[4] (Schapire, 1990; Freund and Schapire, 1996). It is a very nice feature, since the only thing we need to make a stance on is the form of the weak learner.

The goal of boosting is to solve the following optimization problem

$$\min_f \sum_{i=1}^{N} L(y_i, f(x_i)), \tag{1.6}$$

where $L(y,y')$ is a loss function and f is defined by (1.5). Since the boosting estimator depends on the choice of the loss function, the algorithm to solve (1.6) should be adjusted for a particular choice. Instead, one can use a generic version called *gradient boosting* (Friedman, 2001; Mason et al., 2000), which works for an arbitrary loss function.

Breiman (1998) showed that boosting can be interpreted as a form of the gradient descent algorithm in function space. This idea then was further extended by Friedman (2001) who presented the following functional gradient descent or gradient boosting algorithm:

1. Given data $\{(y_i, x_i)\}_{i=1}^n$, initialize the algorithm with some starting value. Common

[4] This is relevant when applied to classification problems. For regression problems any simple method such as least squares regression, regression stump, or one or two-layered neural network will work.

choices are

$$f_0(x) \equiv \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i,c),$$

which is simply $\bar{y}$ under the squared loss, or $f_0(x) \equiv 0$. Set $m = 0$.

2. Increase m by 1. Compute the negative gradient vector and evaluate it at $f_{m-1}(x_i)$:

$$r_{im} = -\left.\frac{\partial L(y_i, f)}{\partial f}\right|_{f=f_{m-1}(x_i)}, \quad i = 1, \ldots, n.$$

3. Use the weak learner to compute (α_m, θ_m) which minimize $\sum_{i=1}^{N}(r_{im} - \alpha\phi(x_i;\theta))^2$.

4. Update

$$f_m(x) = f_{m-1}(x) + \alpha_m \phi(x;\theta_m),$$

that is, proceed along an estimate of the negative gradient vector. In practice, better (test set) performance can be obtained by performing "partial updates" of the form

$$f_m(x) = f_{m-1}(x) + \nu\alpha_m \phi(x;\theta_m),$$

where $0 \leq \nu \leq 1$ is a shrinkage parameter, usually set close to zero (Friedman, 2001).

5. Iterate steps 2 to 4 until $m = M$ for some stopping iteration M.

The key point is that we do not go back and adjust earlier parameters. The resulting basis functions learnt from the data are $\phi(x) = (\phi(x;\theta_1), \ldots, \phi(x;\theta_M))$. The number of iterations M is a tuning parameter, which can be optimally tuned via cross-validation or some model selection criterion (see Section 1.5 for more details).

1.4. Boosting the IV regression

The main complication in the NPIV set-up is that x is potentially endogenous, otherwise learning the structural function via boosting would be straightforward. Moreover, we cannot learn basis functions in the first step and then construct IVs in the second. Dependence of the basis functions for the structural equation on the instruments and vice versa suggests an iterative algorithm.

Before we introduce the algorithm, we need to set up the boosting IV framework first. Com-

bining (1.1) and (1.5) gives

$$y = \alpha_0 + \sum_{m=1}^{M} \alpha_m \varphi(x; \theta_m) + \varepsilon. \tag{1.7}$$

Hence, the conditional expectation of y given z becomes

$$\mathbb{E}[y|z] = \alpha_0 + \sum_{m=1}^{M} \alpha_m \mathbb{E}[\varphi(x; \theta_m)|z]. \tag{1.8}$$

Note that (1.7) and (1.8) closely resemble their standard NPIV counterparts (1.1) and (2.1). The only difference is that the form of basis functions for boosting must be estimated, while for the standard NPIV it has to be ex-ante specified. Since we assume that f can be approximated by an additive basis function model, equation (1.8) is no longer ill-posed. This can be seen as sieve truncation, which is a standard way to regularize the series NPIV estimator, with M being a regularization parameter. Unlike standard boosting, where the goal is to learn $\mathbb{E}[y|x]$, in the presence of endogeneity, we want to learn $\mathbb{E}[y|z]$, implying that in each boosting iteration we have to learn the conditional expectation of the weak learner given the IVs.

To keep things clear and simple, we focus on L_2-boosting which assumes the squared loss function. Bühlmann and Yu (2003) show that L_2-boosting is equivalent to iterative fitting of residuals. In the IV context, it means that at step m the loss has the form

$$L(y, f_{m-1}(x) + \alpha \mathbb{E}[\varphi(x; \theta)|z]) = (r_m - \alpha \mathbb{E}[\varphi(x; \theta)|z])^2,$$

where $r_m \equiv y - f_{m-1}$ is the current residual. Thus, at step m the optimal parameters minimize the loss between the residuals and the conditional expectation of the weak learner given the instruments,

$$(\alpha_m, \theta_m) = \underset{\alpha, \theta}{\operatorname{argmin}} \sum_{i=1}^{N} (r_{im} - \alpha \mathbb{E}[\varphi(x_i; \theta)|z_i])^2. \tag{1.9}$$

However, the conditional expectation $\mathbb{E}[\varphi(x; \theta)|z]$ is unknown and has to be estimated.

A simple way to estimate the conditional expectation in (1.9) is to project[5] the weak learner

[5] In general we do not have to use a projection, we can use a more complex model to estimate the conditional expectation.

on the space spanned by IVs

$$\hat{\mathbb{E}}[\varphi(x;\theta)|z] = P_Z \varphi(x;\theta),$$

where $P_A = A(A'A)^{-1}A'$ is a projection matrix. The exogeneity condition in (1.1) implies that any function of z can serve as an instrument. However, we do not need any function, we need such a transformation of z that will give us strong instruments, i.e. instruments that explain the majority of the variation in the endogenous variables. We follow Gandhi et al. (2019) and introduce an additional step on which we learn the instruments. Let $\mathcal{H}(\cdot;\eta)$ be a class of IV functions parameterized by η. This formulation allows us to use various off-the-shelf algorithms such as Neural Networks, Random Forests, etc. to learn $\mathcal{H}(\cdot;\eta)$. Given the learnt IV transformation $\mathcal{H}(\cdot;\eta)$, we can rewrite (1.9) as

$$(\alpha_m, \theta_m) = \underset{\alpha,\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (r_{im} - \alpha P_{\mathcal{H}(z_i;\eta)} \varphi(x_i;\theta))^2.$$

Since the basis function parameters (α, θ) depend on the IV transformation parameters η and vice versa, we propose an algorithm that iterates between two steps. At the first step we learn instruments, i.e. η_m, given the basis functions parameter estimates from the previous iteration $(\alpha_{m-1}, \theta_{m-1})$, then at the second step we learn new parameter estimates (α_m, θ_m) given the instruments from the first step. We can draw an analogy with the canonical two-stage least squares, where we estimate the reduced form in the first stage, and the structural equation in the second. The details are provided in Algorithm 6.

Algorithm 1 Naive boostIV

Initialize basis functions: $\varphi_0 = \bar{y}$
for iteration m **do**
  First stage: given $\varphi(x;\theta_{m-1})$, estimate $\mathcal{H}(z;\eta_m)$
  Second stage: given $\mathcal{H}(z;\eta_m)$, solve

$$(\alpha_m, \theta_m) = \underset{\alpha,\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(r_{im} - \alpha P_{\mathcal{H}(z_i;\eta_m)} \varphi(x_i;\theta)\right)^2$$

update: $f_m(x) = f_{m-1}(x) + \alpha_m \varphi(x;\theta_m)$
end for
Stop at iteration M

We call this algorithm the Naive boostIV, since we use the same data to learn both the instruments and the basis functions. Asymptotically this will not affect the properties of the estimator, however, in finite samples biases from the first stage will propagate to the

second. This issue can be especially severe if we use regularized estimators in the first stage as the regularization bias will heavily affect the second stage estimates. To get around this issue we resort to cross-fitting.

Let $D = \{y_i, x_i, z_i\}_{i=1}^n$ be our data set, where D_i are iid. Split the data set into a K-fold partition, such that each partition D_k has size $\lfloor \frac{n}{K} \rfloor$, and let D_k^c be the excluded data. The boostIV procedure with cross-fitting is described in Algorithm 2.

Algorithm 2 boostIV with cross-fitting

Folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\} \leftarrow \textsc{Partition}(\mathcal{D}, K)$
Initialize basis functions: $\varphi_0^k = \bar{y}$ for $k = 1, \ldots, K$
for iteration m **do**
  for fold k **do**
    First stage:
      - given $\varphi(x_k^c; \theta_{m-1}^k)$ and z_k^c, estimate $\mathcal{H}(\cdot; \eta_m^k)$
      - apply the learnt transformation to generate IVs $\mathcal{H}(z_k; \eta_m^k)$

    Second stage: Given $\mathcal{H}(z_k; \eta_m^k)$, solve

$$(\alpha_m^k, \theta_m^k) = \underset{\alpha,\theta}{\operatorname{argmin}} \sum_{i \in \mathcal{D}_k} \left(r_{im} - \alpha P_{\mathcal{H}(z_i;\eta_m^k)} \varphi(x_i; \theta)\right)^2$$

    update: $f_m^k(x_k) = f_{m-1}^k(x_k) + \alpha_m^k \varphi(x_k; \theta_m^k)$
  end for
end for
Stop at iteration M: $\hat{f}(x) = \frac{1}{K} \sum_{k=1}^K f_M^k(x)$

1.4.1. *Learning optimal instruments*

Our boostIV algorithm also allows to incorporate optimal instruments in the sense of Chamberlain (1987), i.e. instruments that achieve the smallest asymptotic variance. Assuming conditional homoskedasticity, the optimal instrument vector of Chamberlain (1987) at step m is

$$\mathcal{H}(z; \eta_m) = D_m(z) \sigma_m^{-2}, \tag{1.10}$$

where

$$D_m(z) = \mathbb{E}\left[\frac{\partial \varepsilon(\gamma_m)}{\partial \gamma_m'} \middle| z \right], \quad \gamma_m = (\alpha_m, \theta_m')' \tag{1.11}$$

is the conditional expectation of the derivative of the conditional moment restriction with respect to the boosting parameters, and $\sigma_m^2 = \mathbb{E}[r_m^2 | z]$ is the conditional variance of the error term at step m. Thus, the IV transformation parameters η_m are implicitly embedded in a particular approximation used to estimate $D_m(z)$.

The main complication with using optimal IVs is that they are generally unknown, hence, the common approach is to consider approximations. The parametrization in (1.10)-(1.11)

allows us to use any off-the-shelf statistical/ML method to estimate the optimal functional form for the instruments. Moreover, the iterative nature of the algorithm allows us to use the estimates from step $m-1$ as proxies.

1.4.2. Post-processing

An important feature of the forward stage-wise additive modeling is that we do not go back and adjust earlier parameters. However, we might want to revisit the weights on the learnt basis functions to achieve a better fit. This can be seen as a way of post-processing our boostIV procedure.

The whole procedure can be broken down into two stages:

1. Apply the boostIV algorithm to learn basis functions $\hat{\varphi}_m(x) = \frac{1}{K}\sum_{k=1}^{K}\varphi(x;\theta_m^k)$ for $m = 1, \ldots, M$;

2. Estimate the weights

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{m=1}^{M}\beta_m\hat{\varphi}_m(x_i)\right)^2. \tag{1.12}$$

Note that the basis functions $(\hat{\varphi}_1(x), \ldots, \hat{\varphi}_M(x))$ are causal in the sense that they are constructed using estimated parameters θ that identify a causal relationship between x and y. A more detailed algorithm is presented below.

Algorithm 3 post-boostIV

Folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_L\} \leftarrow \textsc{Partition}(\mathcal{D}, L)$
for fold ℓ **do**
1. apply boostIV to $\mathcal{D}_\ell^c$ and estimate basis functions $(\hat{\varphi}_1^\ell(x), \ldots, \hat{\varphi}_M^\ell(x))$
2. estimate post-boosting weights

$$\hat{\beta}^\ell = \underset{\beta}{\operatorname{argmin}} \sum_{i\in\mathcal{D}_\ell}\left(y_i - \beta_0 - \sum_{m=1}^{M}\beta_m\hat{\varphi}_m^\ell(x_i)\right)^2.$$

3. fold fit at point x: $g^\ell(x) = \hat{\beta}_0^\ell + \sum_{m=1}^{M}\hat{\beta}_m^\ell\hat{\varphi}_m^\ell(x)$

end for
Stop at iteration M: $\hat{f}(x) = \frac{1}{L}\sum_{\ell=1}^{L} f_M^\ell(x)$

Boosting is an example of an ensemble method which combines various predictions with appropriate weights to get a better prediction. In the context of boostIV it works in the following way. We exploit the variation in IVs to get causal parameters θ. Given the esti-

mated parameters, we can treat each learnt basis function $\varphi(x;\theta_m)$ as a separate prediction obtained by fitting a base learner. Then, the post-processing step in (1.12) can be simply seen as model averaging.

We can estimate optimal weights $\hat{\beta}$ by simply running a least squares regression as in (1.12) or use any other method such as Random Forrests, Neural Networks, boosting, etc. To avoid carrying over any biases from the estimation of $(\hat{\varphi}_1(x), \ldots, \hat{\varphi}_M(x))$ into the choice of β, we use cross-fitting once again, which is a generalization of the stacking idea of Wolpert (1992).

1.5. Choosing the optimal number of boosting iterations

Boosting performance crucially depends on the number of boosting iterations, in other words, M is a tuning parameter. A common way to tune any ML algorithm is cross-validation (CV). The most popular type of CV is k-fold CV. The idea behind k-fold CV is to create a number of partitions (validation datasets) from the training dataset and fit the model to the training dataset (sans the validation data). The model is then evaluated against each validation dataset and the results are averaged to obtain the cross-validation error. In application to boosting, we can estimate the CV error for a grid of candidate tuning parameters (number of iterations) and pick M^* that minimizes the CV error. Alternatively, Bühlmann and Hothorn (2007) show how to apply AIC and BIC criteria to boosting in the exogenous case. However, it is not clear how to adjust those criteria for the presence of endogeneity.

Both the standard k-fold cross validation and the model selection criteria considered in Bühlmann and Hothorn (2007) can be computationally costly as it is necessary to compute all boosting iterations under consideration for the training data. To surpass this issue, we apply *early stopping* to k-fold CV. The idea behind early stopping is to monitor the behavior of the CV error and stop as soon as the performance starts decreasing, i.e. CV error goes up.

Algorithm 4 provides implementation details for the k-fold CV with early stopping for either boostIV or post-boostIV procedure. The early stopping criterion compares the CV error evaluated for the model based on M_j boosting iterations to the CV error evaluated for the model based on M_i, $M_i < M_j$. If $CV^{err}(M_j) > CV^{err}(M_i) + \epsilon$, where $\epsilon > 0$ but close to zero is a numerical error tolerance level, then we stop and set $M^* = M_i$, otherwise, continue the search. If the criterion is not met for any of the candidate tuning parameters, we pick the largest value $M^* = \bar{M}$.

An alternative solution would be to use a slice of the dataset as the validation sample and tune the number of iterations using the observations from the validation sample. We actu-

ally use this approach in our simulations since it significantly reduces the computational burden.

Algorithm 4 k-fold CV with early stopping

Folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\} \leftarrow \textsc{Partition}(\mathcal{D}, k)$
Set of indices $\mathcal{I}_M$ corresponding to a sorted grid of tuning parameters $\mathcal{M} = \{1, \ldots, \bar{M}\}$
while $\mathcal{M}[i] \leq \bar{M}$ for $i \in \mathcal{I}_M$ **do**
  for fold $\kappa = 1, \ldots, k$ **do**
    1. training set $\mathcal{T}_\kappa = \mathcal{D}_\kappa^c \rightarrow$ apply (post-)boostIV$(\mathcal{T}_\kappa, \mathcal{M}[i]) \rightarrow f^{boost}_{\mathcal{M}[i],\kappa}(x)$
    2. validation set $\mathcal{V}_\kappa = \mathcal{D}_\kappa \rightarrow CV^{err}_\kappa(\mathcal{M}[i]) = \frac{1}{|\mathcal{V}_\kappa|} \sum_{i \in \mathcal{V}_\kappa} \left(y_i - f^{boost}_{\mathcal{M}[i],\kappa}(x_i)\right)^2$
  end for
  Calculate $CV^{err}(\mathcal{M}[i]) = \frac{1}{k} \sum_{\kappa=1}^{k} CV^{err}_\kappa(\mathcal{M}[i])$
  if $CV^{err}(\mathcal{M}[i]) > CV^{err}(\mathcal{M}[i-1]) + \epsilon$ **then**
    $M^* = \mathcal{M}[i-1]$
  else
    $i = +1$
  end if
end while
$M^* = \bar{M}$
$f^{boost}_{M^*}(x) \leftarrow$ (post-)boostIV$(\mathcal{D}, M^*)$

1.6. Theoretical properties

In this section, we show that under mild conditions boostIV is consistent. Theoretical properties of post-boostIV are beyond the scope of the paper and are left for future research.

We borrow the main idea from Zhang and Yu (2005) and modify it accordingly to apply it to the GMM criterion. Let $g(W_i, f) = (y_i - f(x_i))z_i$ denote a $k \times 1$ moment function, then $g_0(f) = \mathbb{E}[g(W_i, f)]$ is the population moment function and $\hat{g}(f) = n^{-1} \sum_{i=1}^{n} g(W_i, f)$ is its sample analog. Also let Ω denote a $k \times k$ positive semi-definite weight matrix and $\hat{\Omega}$ be its sample analog. Thus, the population GMM criterion and its sample analog are

$$Q(f) = g_0(f)'\Omega g_0(f), \quad \hat{Q}(f) = \hat{g}(f)'\hat{\Omega}\hat{g}(f). \tag{1.13}$$

The form of the GMM criterion in (1.13) corresponds to the form of the empirical objective function in Zhang and Yu (2005) with the loss function replaced by the moment function.

We follow Zhang and Yu (2005) and replace the functional gradient decent step (1.9) leading to the 2SLS fitting procedure on every iteration with an approximate minimization involving a GMM criterion. We can do that since the 2SLS solution is a special case of a GMM solution with an appropriate weighting matrix.

Assumption 1. Approximate Minimization. On each iteration step m we find $\bar{\alpha}_m \in \Lambda_m$ and $\bar{\varphi}_m \in \mathcal{S}$ such that

$$\hat{Q}(f_m + \bar{\alpha}_m \bar{\varphi}_m) \leq \inf_{\alpha_m \in \Lambda_m \varphi_m \in \mathcal{S}} \hat{Q}(f_m + \alpha_m \varphi_m) + \epsilon_m, \tag{1.14}$$

where ϵ_m is a sequence of non-negative numbers that converge to 0.

As Zhang and Yu (2005) show, the consistency of the boosting procedure consists of two parts: (i) numerical convergence of the procedure itself, i.e. the algorithm achieves the true minimum of the objective function, and (ii) statistical convergence that ensures the uniform convergence of the sample criterion to its population analog. We will treat these two steps separately in the following subsections, and then combine them to demonstrate consistency of the boostIV.

1.6.1. Numerical Convergence

To demonstrate numerical convergence, we first have to verify that the sample GMM criterion in (1.13) satisfies Assumption 3.1 from Zhang and Yu (2005).

Following Zhang and Yu (2005), we introduce some additional notation. Let $\mathcal{S}$ be a set of real-valued functions and define

$$\text{span}(\mathcal{S}) = \left\{ \sum_{j=1}^{J} w_j \varphi_j : \varphi_j \in \mathcal{S}, w_i \in \mathbb{R}, J \in \mathbb{Z}^+ \right\},$$

which forms a linear function space. Also, for all $f \in \text{span}(S)$ define the 1-norm with respect to the basis S as

$$||f||_1 = \left\{ ||w||_1 : f = \sum_{j=1}^{J} w_j \varphi_j : \varphi_j \in \mathcal{S}, J \in \mathbb{Z}^+ \right\}.$$

Assumption 2. A convex function $A(f)$ defined on $\text{span}(\mathcal{S})$ should satisfy the following conditions:

1. The functional A satisfies the following Frechet-like differentiability condition

$$\lim_{h \to 0} \frac{1}{h}(A(f + h\varphi) - A(f)) = \nabla A' \varphi$$

2. For all $f \in \text{span}(\mathcal{S})$ and $\varphi \in \mathcal{S}$, the real-valued function $A_{f,\varphi}(h) = A(f + h\varphi)$ is

second-order differentiable (as a function of h) and the second derivative satisfies

$$A''_{f,\varphi}(0) \leq M(||f||_1),$$

where $M(\cdot)$ is a nondecreasing real-valued function.

Lemma 1. Let (i) the basis functions φ be bounded as $\sup_x |\varphi(x)^2| = C < \infty$, (ii) the maximal eigenvalue λ_{max} of the weighting matrix Ω be bounded from above, $\lambda_{max}(\Omega) < \infty$, and (iii) $\mathbb{E}[|z_i' z_i|] \leq B < \infty$. Then the population GMM criterion defined in (1.13) satisfies Assumption 2.

Assumption 3. Step size.

(a) Let $\Lambda_m \subset \mathbb{R}$ such that $0 \in \Lambda_m$ and $\Lambda_m = -\Lambda_m$.

(b) Let $h_m = \sup \Lambda_m$ satisfy the conditions

$$\sum_{j=0}^{\infty} h_j = \infty, \quad \sum_{j=0}^{\infty} h_j^2 < \infty. \tag{1.15}$$

Then we can bound the step size $|\bar{\alpha}_m| \leq h_m$.

Note that Assumption 3(a) restricts the step size α_m. Friedman (2001) argues that restricting the step size is always preferable in practice, thus, we will restrict our attention to this case[6]. Moreover, Λ_m is allowed to depend on the previous steps of the algorithm. Assumption 3(b) requires the step size h_j to be small ($\sum_{j=0}^{\infty} h_j^2 < \infty$) preventing large oscillation, but not too small ($\sum_{j=0}^{\infty} h_j = \infty$) ensuring that f_m can cover the whole span($\mathcal{S}$). The following theorem establishes the main numerical convergence result.

Theorem 1. Assume that we choose quantities f_0, ϵ_m and Λ_m independent of the sample W. Given the results of Lemma 1, as long as there exists h_j satisfying Assumption 3 and ϵ_j such that $\sum_{j=0}^{\infty} \epsilon_j < \infty$, we have the following convergence result:

$$\lim_{m \to \infty} \hat{Q}(f_m) = \inf_{f \in \text{span}(\mathcal{S})} \hat{Q}(f).$$

[6] Zhang and Yu (2005) provide a short discussion on how to deal with the unrestricted step size, however, the argument relies on the exact minimization which greatly complicates the analysis.

1.6.2. Statistical convergence

We need to show that the sample GMM criterion uniformly converges to its population analog, then under proper regularity conditions we will be able to ensure consistency of boostIV.

To show that the sample GMM criterion converges uniformly to its population analog, we will first bound the moment function and then we will show that it is sufficient to put a bound on the criterion function.

Assumption 4. Assume the following conditions:

1. The class of weak learners $\mathcal{S}$ is closed under negation, i.e. $f \in \mathcal{S} \to -f \in \mathcal{S}$.

2. The moment function is Lipschitz with each component $j = 1, \ldots, k$ satisfying

$$\exists \gamma_j(\beta) \text{ such that } \forall |f_1|, |f_2| \leq \beta \quad |g_j(f_1) - g_j(f_2)| \leq \gamma_j(\beta)|f_1 - f_2|,$$

   implying that

$$||g(f_1) - g(f_2)|| \leq \gamma(\beta)|f_1 - f_2|, \quad \gamma(\beta) = \sqrt{\sum_{j=1}^{k} \gamma_j^2(\beta)}.$$

To bound the rate of uniform convergence of the moment function, we appeal to the concept of Rademacher complexity. Let $\mathcal{H} = h(w)$ be a set of real-valued functions. Let $\{\zeta_i\}_{i=1}^n$ be a sequence of binary random variables such that ζ_i takes values in $\{-1,1\}$ with equal probabilities. Then the sample or empirical Rademacher complexity of class $\mathcal{H}$ is given by

$$\hat{R}(\mathcal{H}) = \mathbb{E}_\zeta \left[\sup_{h \in \mathcal{H}} n^{-1} \sum_{i=1}^{n} \zeta_i h(W_i) \right]. \tag{1.16}$$

We also denote $R(\mathcal{H}) = \mathbb{E}_W \hat{R}(\mathcal{H})$ to be the expected Rademacher complexity, where $\mathbb{E}_W$ is the expectation with respect to the sample $W = (W_1, \ldots, W_n)$. Note that the definition in (1.16) differs from the standard definition of Rademacher complexity where there is an absolute value under the supremum sign (Vaart and Wellner, 1996). The current version of Rademacher complexity has the merit that it vanishes for function classes consisting of single constant function, and is always dominated by the standard Rademacher complexity. Both definitions agree for function classes which are closed under negation (Meir and Zhang, 2003).

Lemma 2. Under Assumption 4, for all $j = 1, \ldots, k$,

$$\mathbb{E}_W \sup_{||f||_1 \leq \beta} |g_{0,j}(f) - \hat{g}_j(f)| \leq 2\gamma_j(\beta)\beta R(\mathcal{S}).$$

Note that β controls the complexity of f with respect to the span of $\mathcal{S}$. The more base learners in the approximation, the harder is the target function to learn.

For many classes the Rademacher complexity can be calculated directly, however, to obtain a more general result we need to bound $R(\mathcal{S})$. Using the results from Section 4.3 in Zhang and Yu (2005) we can bound the expected Rademacher complexity of the weak learner class by

$$R(\mathcal{S}) \leq \frac{C_{\mathcal{S}}}{\sqrt{n}}, \tag{1.17}$$

where $C_{\mathcal{S}}$ is a constant that solely depends on $\mathcal{S}$. Zhang and Yu (2005) also show that popular weak learners such as two-level neural networks and trees basis functions satisfy the requirements. However, Zhang and Yu (2005) point out that in general the bound may be slower than root-n. In Appendix A.3 we derive an alternative bound on $R(\mathcal{S})$ that works for any class with finite VC dimension. The derived VC bound is slower by the factor of $\log(n)$ that appears in a lot of ML algorithms.

Condition (1.17) allows us to bound the moment function which leads to a bound on the rate of uniform convergence of the GMM criterion. The formal statements of the results are presented below.

Lemma 3. Suppose that condition (1.17) holds, then under Assumption 4,

$$\sup_{||f||_1 \leq \beta} ||g_0(f) - \hat{g}(f)|| \xrightarrow{p} 0.$$

Theorem 2. Suppose that (i) the data $W = (W_1, \ldots, W_n)$ are i.i.d., (ii) $\hat{\Omega} \xrightarrow{p} \Omega$, (iii) Assumption 4 is satisfied, and (iv) $\mathbb{E}_W \left[\sup_{||f||_1 \leq \beta} ||g(W_i, f)||\right] < \infty$. Then

$$\sup_{||f||_1 \leq \beta} |\hat{Q}(f) - Q(f)| \xrightarrow{p} 0.$$

1.6.3. Consistency

In this section we put together the arguments for numerical and statistical convergence presented in the previous subsections to prove consistency of the boostIV algorithm. We

start with a general decomposition illustrating the proof strategy and highlighting where exactly numerical and statistical convergence step in.

Suppose that we run the boostIV algorithm and stop at an early stopping point $\hat{m}$ that satisfies $\mathbb{P}(||\hat{f}_{\hat{m}}||_1 \leq \beta) = 1$ for some sample-independent $\beta \geq 0$. Let f^* be a unique minimizer of the population criterion, i.e. $Q(f^*) = \inf_{f \in \text{span}(\mathcal{S})} Q(f)$. By the triangle inequality, we get the following decomposition

$$\begin{aligned}\left|Q(\hat{f}_{\hat{m}}) - Q(f^*)\right| &\leq \left|Q(\hat{f}_{\hat{m}}) - \hat{Q}(\hat{f}_{\hat{m}})\right| + \left|\hat{Q}(\hat{f}_{\hat{m}}) - \hat{Q}(f^*)\right| + \left|\hat{Q}(f^*) - Q(f^*)\right| \\ &\leq 2 \sup_{||f||_1 \leq \beta} \left|\hat{Q}(f) - Q(f)\right| + \left|\hat{Q}(\hat{f}_{\hat{m}}) - \hat{Q}(f^*)\right|\end{aligned}$$

We can bound the first term using the uniform bound on the sample GMM criterion in Theorem 2, this is the statistical convergence argument. In order to bound the second term, we have to appeal to the numerical convergence argument in Theorem 1. As a result, since $Q(\hat{f}_{\hat{m}}) \to Q(f^*)$ as $n \to \infty$, it follows that $\hat{f}_{\hat{m}} \xrightarrow{p} f^*$. The following theorem formalizes the result.

Theorem 3. Suppose that the assumptions of Theorems 1 and 2 hold. Consider two sequences k_n and β_n such that $\lim_{n\to\infty} m_n = \infty$ and $\lim_{n\to\infty} \gamma(\beta_n)\beta_n R(\mathcal{S}) = 0$. Then as long as we stop the algorithm at step $\hat{m}$ based on W such that $\hat{m} \geq m_n$ and $||\hat{f}_{\hat{m}}||_1 \leq \beta_n$, we have the consistency result $\hat{f}_{\hat{m}} \xrightarrow{p} f^*$.

Note that in Theorem 3 we allow for β_n to grow with the sample size. In other words, more data allows us to learn a more complex function with the desired level of generalization.

1.7. Monte Carlo experiments

1.7.1. Univariate design

To begin with, we consider a simple low-dimensional scenario with one endogenous variable and two instruments.

$$y = g(x) + \rho e + \delta, \quad x = z_1 + z_2 + e + \gamma,$$

where instruments $z_j \sim U[-3, 3]$ for $j = 1, 2$, $e \sim \mathcal{N}(0, 1)$ is the confounder, $\delta, \gamma \sim \mathcal{N}(0, 0.1)$ are additional noise components, and ρ is the parameter measuring the degree of endogeneity, which we set to 0.5 in the simulations. We focus on four specifications of the structural function:

- **abs:** $g(x) = |x|$

- **log:** $g(x) = \log(|16x - 8| + 1)\text{sign}(x - 0.5)$
- **sin:** $g(x) = \sin(x)$
- **step:** $g(x) = \mathbb{1}\{x < 0\} + 2.5 \times \mathbb{1}\{x \geq 0\}$

We compare the performance of boostIV and post-boostIV with the standard NPIV estimator using the cubic polynomial basis, Kernel IV (KIV) regression of Singh et al. (2019)[7], DeepIV estimator of Hartford et al. (2017*a*)[8] and DeepGMM estimator of Bennett et al. (2019)[9]. We use $1{,}000$ observations for both train and test sets and 500 observations for the validation set. Our results are based on 200 simulations for each scenario.

Table 1: Univariate design: Out-of-sample MSE.

	NPIV	KIV	DeepIV	DeepGMM	boostIV	post-boostIV
abs	0.1916	0.0564	0.1347	1.2717	0.0348	0.0217
log	0.6936	0.3367	1.2708	14.4615	0.3173	0.0930
sin	0.1837	0.0217	0.2798	0.8595	0.0292	0.0124
step	0.1267	0.0972	0.1756	0.9796	0.1027	0.0546

We plot our results in Figure 1 which shows the average out of sample fit across simulations (orange line) compared to the true target function (black line). Table 1 presents the out-of-sample MSE across simulations. First thing to notice is that NPIV fails to capture different functional form subtleties. Second, DeepIV's performance does not improve upon the one of NPIV. Moreover, even though DeepGMM estimates have lower bias than the ones of NPIV and DeepIV (except for the log function), they are quite volatile across simulations leading to higher MSE. BoostIV performs on par with KIV both in terms of the bias term as they are able to recover the underlying structural relation, and in terms of the variance leading to low MSE. Finally, the post-processing step helps to further improve upon boostIV's performance by reducing bias. On top of that, post-boostIV requires less iterations to converge. We use $5{,}000$ iterations for boostIV, while post-boostIV uses on average 50 iterations.[10]

[7]Code: https://github.com/r4hu1-5in9h/KIV

[8]We use the latest implementation of the econML package: https://github.com/microsoft/EconML

[9]Code: https://github.com/CausalML/DeepGMM

[10]In this experiment we do not tune boostIV, we just pick a large enough number of iterations for it to converge. However, we do tune post-boostIV.

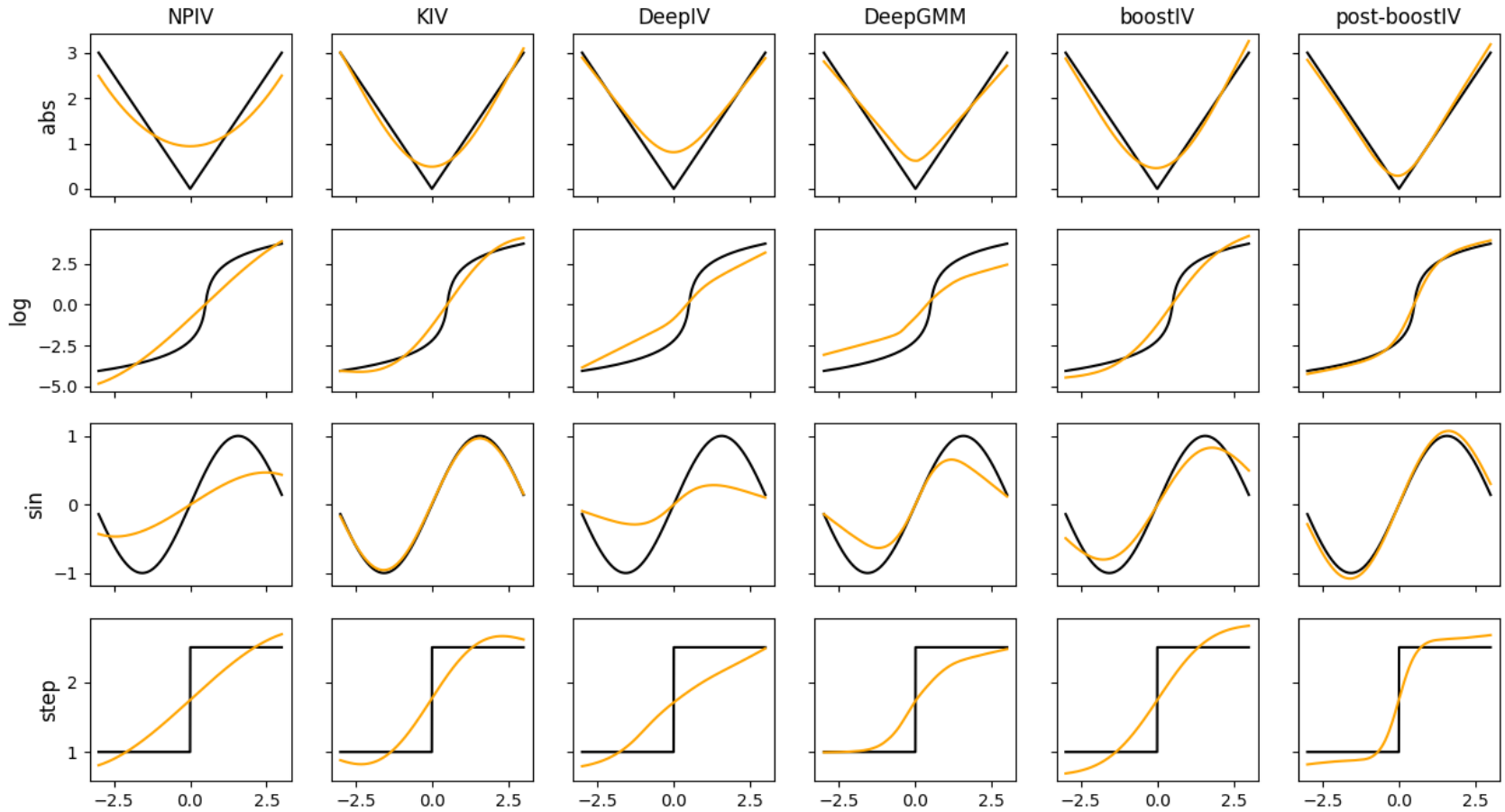

Figure 1: Out-of-sample average fit across simulations.

1.7.2. Multivariate Design

Consider the following data generating process:

$$y_i = h(x_i) + \varepsilon_i$$
$$x_{i,k} = g_k(z_i) + v_{i,k}, \quad k = 1, \ldots, d_x,$$

where $y_i \in \mathbb{R}$ is the response variable, $x_i \in \mathbb{R}^{d_x}$ is the vector of potentially endogenous variables, $z_i \in \mathbb{R}^{d_z}$ is the vector of instruments, $\varepsilon_i \in \mathbb{R}$ is the structural error term, and $v_i \in \mathbb{R}^{d_x}$ is the vector of the reduced form errors. Function $h(\cdot)$ is the structural function of interest, and function $g(\cdot)$ governs the reduced form relationship between the endogenous regressors and instrumental variables.

Instruments are drawn from a multivariate normal distribution, $z_i \sim \mathcal{N}(0, \Sigma_z)$, where Σ_z is just an identity matrix. The error terms are described by the following relationship:

$$\varepsilon \sim \mathcal{N}(0, 1), \quad v \sim \mathcal{N}(\rho\varepsilon, \mathcal{I} - \rho^2),$$

where ρ is the correlation between ε and the elements of v, which controls the degree of endogeneity.

We consider two structural function specifications:

1. a simpler design where the structural function is proportional to a multivariate normal density, i.e. $h(x) = \exp\{-0.5x'x\}$. We will further refer to this specification as Design 1;

2. a more challenging design where the structural function is $h(x) = \sum_{k=1}^{d_x} \sin(10x_k)$. We will further refer to this specification as Design 2.

We also consider two different choices of the reduced form function $g(\cdot)$:

(a) linear: $g(Z_i) = Z_i'\Pi$, where $\Pi \in \mathbb{R}^{d_z \times 1}$ is a matrix of reduced form parameters;

(b) non-linear: $g_k(Z_i) = G(Z_i; \theta_k)$ for $k = 1, \ldots, d_x$, where $G(Z_i; \theta_k)$ is a multivariate normal density parameterized by the mean vector θ_k (for simplicity, we use the identity covariance matrix).

We use $1{,}000$ observations for the train set and 500 observations for both the validation and test sets. We run 200 simulations for each scenario. The results are summarized in Tables 2 and 3. We observe that boostIV performs on par with KIV and DeepIV and slightly outperforms DeepGMM. Unlike the univariate case, post-boostIV does not improve upon

the boostIV, but it still significantly outperforms NPIV.

Table 2: Design 1: Out-of-sample MSE.

dx	dz	IV type	ρ	NPIV	KIV	DeepIV	DeepGMM	boostIV	post-boostIV
5	7	lin	0.25	4.9535	0.0147	0.0497	0.2234	0.0213	1.3306
			0.75	6.8889	0.0286	0.054	0.1655	0.0603	0.7249
		nonlin	0.25	4.0548	0.017	0.0757	0.3262	0.0875	0.3457
			0.75	1.9932	0.0516	0.1188	0.7265	0.4287	0.9128
10	12	lin	0.25	23.1025	0.0024	0.0867	0.3089	0.0084	1.0953
			0.75	39.6902	0.0108	0.0884	0.251	0.0427	0.6347
		nonlin	0.25	6.4842	0.0038	0.05	0.4908	0.0525	0.8137
			0.75	2.53	0.0147	0.0691	0.822	0.3332	0.8937

Table 3: Design 2: Out-of-sample MSE.

dx	dz	IV type	ρ	NPIV	KIV	DeepIV	DeepGMM	boostIV	post-boostIV
5	7	lin	0.25	21.5854	2.4983	2.5484	2.9105	2.5105	3.5498
			0.75	23.2413	2.5043	2.5358	2.792	2.5351	3.5081
		nonlin	0.25	19.0871	2.5118	2.5415	2.9867	2.5707	3.0303
			0.75	22.1192	2.5367	2.5523	3.4188	2.891	3.7043
10	12	lin	0.25	147.984	5.0047	5.1383	5.9647	5.0209	5.9435
			0.75	241.56	5.0326	5.1698	5.7147	5.0781	6.3259
		nonlin	0.25	61.0328	5.0103	5.0636	6.2145	5.0713	5.9001
			0.75	112.785	4.9799	5.0631	6.193	5.3172	6.4674

1.8. Conclusion

In this Chapter we have introduced a new boosting algorithm called boostIV that allows to learn the target function in the presence of endogenous regressors. The algorithm is very intuitive as it resembles an iterative version of the standard 2SLS regression. We also study several extensions including the use of optimal instruments and the post-processing step.

We show that boostIV is consistent and demonstrates an outstanding finite sample performance in the series of Monte Carlo experiments. It performs especially well in the non-parameteric demand estimation example which is characterized by a complex nonlinear relationship between the target function and explanatory features.

Despite all the advantages of boostIV, the algorithm does not allow for high-dimensional settings where the number of regressors and/or instruments exceeds the number of observations. We also believe it is possible to extend our algorithm in the spirit similar to XGBoost (Chen and Guestrin, 2016) that could decrease the computation time taken by the algorithm. These would be interesting directions for future research.

CHAPTER 2: Automatic Debiased Machine Learning in Presence of Endogeneity

2.1. Introduction

Instrumental variables methods are widely used in applied research for estimation and inference in models containing endogenous regressors. In many cases, economic theory does not impose any functional form restrictions motivating nonparametric instrumental variables (NPIV) methods, where the function of interest is not assumed to be known up to a finite-dimensional parameter. In many cases, structural parameters of economic interest appear as functionals of that underlying unknown function. Examples are policy effects, average (weighted) partial effects, consumer surplus, measures of substitution patterns, and various counterfactuals from structural models. It is quite common for the estimation problem to be high-dimensional. There might be many control variables which we want to include in a flexible way along with the endogenous regressor, or a structural model may depend on many variables, e.g. in the demand for differentiated goods framework, the demand function depends on the vector of prices and product characteristics of all products in the market. In this paper, we are interested in estimation and inference on structural economic objects in presence of endogeneity when the dimensionality of the problem is (moderately) high.

Machine learning (ML) literature provides a collection of modern statistical tools for flexible estimation of various statistical objects that are especially powerful in high-dimensional settings. However, standard ML estimators, such as Lasso, boosting, or Neural Networks are unable to pick up causal relationships when endogenous regressors are present (see e.g., Hartford et al., 2017*b*). On the other hand, there is a new line of research in machine learning and computer science communities that offers a series of new algorithms that both addresses endogeneity and can be applied in high-dimensional environments, we refer to them as MLIV estimators. These algorithms are data-driven and exploit various forms of regularization to ameliorate the ill-posedness of the problem while maintaining the functional form flexibility. Examples include the DeepIV estimator (Hartford et al., 2017*b*), the Kernel IV regression (Singh et al., 2019), the Dual IV regression (Muandet et al., 2019), the DeepGMM estimator (Bennett et al., 2019), the Double Lasso estimator of Gold et al. (2020), a series of estimators constructed using the minimax framework of Dikkala et al. (2020), and the boostIV estimator (Bakhitov and Singh, 2021). The goal of this paper is to use these novel methods to estimate and perform inference on various economic objects of interest that appear as functionals of the underlying structural function under endogeneity.

As standard ML algorithms, MLIV estimators produce inherently biased estimates. The main source of bias is regularization and/or model selection needed to balance out squared bias and variance to obtain overall small mean squared errors. In the NPIV context, regularization is particularly important as it plays a dual role. First, it allows to deal with the curse of dimensionality, as in the case of standard ML estimators. Second, it is necessary to solve the ill-posed problem. As a result, regularization and/or model selection bias leads to poor coverage unless it is corrected for. Furthermore, the bias term will propagate into the functional estimate if we simply plug-in an MLIV estimator into the functional formula. As Chernozhukov, Newey and Robins (2018) point out, squared bias of plug-in estimators can shrink slower than the variance, leading to extremely poor confidence interval coverage.

In this paper, we provide an approach for performing valid asymptotic inference on functionals of MLIV estimators. Our method bases off of the automatic debiased machine learning approach of Chernozhukov, Newey and Singh (2018), hereafter CNS, but focuses on the endogenous setting rather than the exogenous one. To get rid of the regularization and/or model selection bias, we debias the moment function identifying the functional of interest. The debiasing is automatic in the sense that it only depends on the form of the identifying moment function but not on the form of the bias correction term. The key to bias correction is Neyman orthogonality of the moment function which ensures that the estimated moment function has a zero derivative with respect to the MLIV estimator. Intuitively it means that the estimated moment function is insensitive to local perturbations around the true value of the estimated function, which allows to plug-in noisy estimates in the moment condition without strongly violating it. We construct Neyman orthogonal, or simply debiased, moment functions by adding the influence function for the MLIV estimator to the identifying moment functions. Then we simply plug-in the MLIV estimator in the debiased moment function to get the debiased estimate of the functional of interest.

We focus our attention on regular functionals with the finite semiparametric asymptotic variance bound necessary for root-n estimability. We allow for both linear and non-linear functionals, though the conditions for root-n rate are much tighter for the nonlinear case. When the semiparametric asymptotic variance bound is finite, the influence function adjustment term depends on the Riesz representer (RR) for the identifying moment function in case of a linear functional or the derivative of the identifying moment condition in case of a nonlinear functional. Typically, in the NPIV framework, the form of the RR is either very complicated to derive or even unknown. We exploit the orthogonality of the identifying moment condition and provide a penalized GMM (PGMM) framework to estimate the RR. This allows us to learn the RR directly from the identifying moment conditions without requiring the knowledge of the form of the RR, hence, we refer to this estimator

as automatic. The PGMM estimator of the RR is novel and, to the best of our knowledge, is the only automatic estimator of the RR in the NPIV framework. The PGMM estimator is a generalization of the Lasso minimum distance estimator of CNS as it allows for a more general form of the influence function.

We derive the convergence rate for the PGMM estimator and provide conditions for root-n consistency and asymptotic normality of the debiased MLIV estimator of the functional of interest. To accommodate for a large variety of MLIV estimators, we only require certain mean square consistency and convergence rates for MLIV estimators. The required conditions differ quite drastically for linear and nonlinear functionals. For linear functionals it is sufficient to require the MLIV estimator to converge at some positive rate in the projected mean square norm. It is well-known that NPIV estimators exhibit much faster convergence rates in the projected norm rather than in the standard mean square norm due to ill-posedness (see e.g., Blundell et al., 2007; Chen and Pouzo, 2012, 2015). However, for nonlinear functionals it is necessary to account for the linearization bias which requires the convergence rate to be faster than $n^{-1/4}$, which is a standard condition in the semiparametric literature (Newey, 1994). Moreover, the presence of nonlinearities in the identifying moment function results in the convergence rate condition in the standard mean square norm rather than the projected norm, which makes it harder to satisfy in practice.

We apply our approach to learning the conditional demand derivative functional in the nonparametric demand for differentiated products framework (Berry and Haile, 2016; Compiani, 2018; Gandhi et al., 2020) that has been gaining popularity in the last years as an alternative to the standard parametric procedure of Berry et al. (1995), hereafter, BLP. The conditional demand derivative with respect to own price has a nice economic interpretation which has a close connection with traditional parametric models such as logit and nested logit. Under logit, the conditional demand derivative becomes just the logit price coefficient, while under nested logit the derivative consists of two parts: (i) the direct effect from the price coefficient and (ii) the indirect effect coming from the nesting structure. We use these insights and run Monte Carlo experiments where we nonparametrically estimate the conditional demand derivative under logit and nested logit data generating processes. We show that the plug-in estimates are badly biased and have extremely poor coverage as a result. Furthermore, we demonstrate that our debiasing procedure not only significantly reduces bias, but also achieves close to the nominal level coverage.

We use the Monte Carlo results as a basis for our empirical application where we estimate the conditional demand derivative using scanner data. First, we demonstrate that applying machine learning allows to uncover more complicated substitution patterns compared to traditional parametric estimators. The nested logit estimates of the conditional demand

derivative do not exhibit much variation across products and are close the logit price coefficient estimate. While, ML estimates have substantial variation across products and state that similar products have similar responses to price changes. Moreover, our empirical results are coherent with the evidence from the Monte Carlo experiments: plug-in estimates are biased upwards and have smaller standard errors compared to the debiased estimates.

This Chapter connects several strands of literature. First, since the focus of the Chapter is functionals of nonparametric quantities, our methodology relates to the literature on semiparametric statistical theory (Van der Vaart, 1991; Bickel et al., 1993; Newey, 1994; Robins and Rotnitzky, 1995; Van der Vaart, 2000). These papers focus on functionals of densities or regressions in low dimensional settings, while in this Chapter we focus on functionals of MLIV estimators over domains that may include low, moderate, and high dimensional objects. A more recent work by Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) generalizes and extends the insights from the classical theory by constructing Neyman orthogonal moment conditions allowing for a wide range of ML estimators[11]. We follow Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) and use Neyman orthogonal moment functions with the influence function adjustment term for the NPIV estimator from Ichimura and Newey (2017).

Riesz representers are important objects in semiparametric theory as they appear in calculations of the asymptotic variance of functionals of nonparametric quantities (Ichimura and Newey, 2017; Chernozhukov, Escanciano, Ichimura, Newey and Robins, 2020). For the same reason they appear in the influence function calculations, which makes estimation of RRs a cornerstone of the debiased machine learning literature. Chernozhukov et al. (2019) and CNS propose Lasso and Dantzig minimum distance estimators of the RR based on the sparse approximation assumption. While the latter provides asymptotic results for regular functionals, the former provides finite sample analysis and also allows for irregular functionals. A recent paper by Chernozhukov et al. (2021) proposes to use a neural network to estimate the RR. On the other hand, Chernozhukov, Newey, Singh and Syrgkanis (2020) take a different approach and allow for a more general estimator of the RR based on the minimax framework of Dikkala et al. (2020). While the aforementioned papers can be applied only in exogenous settings, the PGMM estimator we propose allows to estimate the RR under endogeneity.

This work also contributes to the literature on estimation and inference on conditional re-

[11] Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) provide high-level conditions for inference on functionals for conditional moment restriction models that nest the NPIV problem (see Theorem 19). Our results are complementary as we provide an estimator of the RR and give low-level conditions to derive its convergence rate.

strictions models which nest the NPIV regression problem as a special case. Several NPIV estimators are now available including kernel-based estimators (Hall and Horowitz, 2005; Darolles et al., 2011) and series or sieve estimators (Newey and Powell, 2003; Blundell et al., 2007; Chen and Pouzo, 2012; Chen et al., 2021). There are several papers focusing on linear regular functionals of NPIV estimators, see e.g., Ai and Chen (2003), Santos (2011), and Severini and Tripathi (2012) among others. Chen and Pouzo (2015) and Chen and Christensen (2018) give conditions for pointwise and uniform asymptotic normality, respectively, of possibly nonlinear functionals of the sieve NPIV estimator. The results presented in this Chapter are complementary to the results on inference on functionals of NPIV estimators.

The remainder of the Chapter is organized as follows. Section 2.2 briefly introduces the NPIV framework, discusses practical issues, and describes various MLIV estimators. In Section 2.3 we describe the objects of interest and provide several economic examples. We also illustrate how to construct the debiased estimator and the estimator of its asymptotic variance. Finally, we introduce the PGMM estimator of the RR. Section 2.4 gives conditions necessary to derive a convergence rate for the PGMM estimator. Section 2.5 gives conditions for root-n consistency and asymptotic normality of the debiased estimator for linear functionals. In Section 2.6 we introduce additional conditions necessary to extend our results to nonlinear functionals. Section 2.7 examines the performance of the debiased estimator in a simple Monte Carlo exercise. Section 2.8 concludes. All additional details and proofs are left for the Appendix.

Notation: For a vector $x \in \mathbb{R}^n$, let $|x|_1$, $||x||$, and $||x||_\infty$ denote its ℓ_1-, ℓ_2-, and ℓ_∞-norms respectively. For an $m \times n$ matrix A, we define $||A||_\infty = \max_{j,k} |A_{jk}|$. Let $||A||_{\ell_\infty} = \max_i \sum_{j=1}^n |A_{ij}|$ denote the induced ℓ_∞-norm of A. For $S \subseteq \{1, \ldots, n\}$ let x_S be the modification of x that places zeros in all entries of x whose index does not belong to S. For a random variable X, let $L_2(X)$ denote a space of all measurable and square integrable functions.

2.2. Flexible estimation under endogeneity

We start with a brief discussion of the NPIV framework and consequences of ill-posedness of the NPIV problem for practitioners and then we categorize and describe various MLIV algorithms.

2.2.1. *Nonparametric IV framework*

Consider the nonparametric instrumental variables framework of Newey and Powell (2003),

$$Y = \gamma_0(X) + \varepsilon, \quad \mathbb{E}[\varepsilon|Z] = 0,$$

where Y is an explanatory variable, X is a vector of potentially endogenous regressors, Z is a vector of instruments, and ε is an error term. Suppose that γ_0 is identified and the completion condition holds, i.e. for all measurable real functions δ with finite expectation,

$$\mathbb{E}[\delta(X)|Z] = 0 \Rightarrow \delta(X) = 0.$$

Intuitively, this condition implies that there is enough variation in the instruments to explain the variation in the endogenous covariates. For example, in the linear model the completeness condition is equivalent to the usual rank condition.

The unknown function γ_0 solves the following integral equation,

$$\mathbb{E}[Y|Z] = \int \gamma_0(x) f(x|z) dx, \tag{2.1}$$

where f denotes the conditional pdf of X given Z. Solving for γ directly is an ill-posed problem as it involves inverting linear compact operators (see e.g., Kress, 1989). Ill-posedness implies that the solution to (2.1) is not continuous in $\mathbb{E}[Y|Z]$ and $f(x|z)$. This leads to certain estimation issues as one cannot construct an estimator of γ by plugging in consistent estimators of $\mathbb{E}[Y|Z]$ and $f(x|z)$ and approximately solving for γ.

A well-known solution to the ill-posed inverse problem is regularization, which means constructing an estimator of γ_0 in a way that ill-posedness does not affect consistency. In essence, regularization allows us to avoid estimation of higher-order terms that drive up the variance. There are several traditional ways to regularize a solution to (2.1). For example, Kress (1989) proposes a very intuitive form of regularization where γ_0 is replaced with a finite dimensional approximation. Another popular method is to use Tikhonov regularization (see e.g., Hall and Horowitz, 2005; Carrasco et al., 2007).

Ill-posedness negatively affects convergence rates of the NPIV estimators making them slower than of the standard nonparametric regression counterparts. To illustrate the issue, we appeal to an important quantity called the measure of ill-posedness which measures how much the conditional expectation in (2.1) smoothes out γ. Let $T : L_2(X) \mapsto L_2(Z)$

denote the conditional expectation operator given by

$$T\gamma = \mathbb{E}[\gamma(X)|Z].$$

Let τ denote the measure of ill-posedness defined as

$$\tau = \sup_{\gamma \in \Gamma} \frac{||\gamma - \gamma_0||}{||T(\gamma - \gamma_0)||},$$

where $\Gamma \subseteq L_2(X)$ and $||T(\gamma - \gamma_0)|| = \sqrt{\mathbb{E}\{\mathbb{E}[\gamma - \gamma_0|Z]\}^2}$ is the projected mean square norm. Typically, τ grows with n, but for simplicity assume that τ is bounded, then

$$||\gamma - \gamma_0|| \leq \tau ||T(\gamma - \gamma_0)||.$$

Thus, the convergence rate in the mean square norm is always slower than the convergence rate in the projected norm. On the other hand, it is possible to obtain fast rates in the projected norm even when the mean square rate is slow as its definition sidesteps ill-posedness (see e.g., Blundell et al., 2007; Chen and Pouzo, 2012; Dikkala et al., 2020).

2.2.2. Practical concerns

Standard NPIV methods provide flexible and intuitive approaches to nonparametric estimation under endogeneity. However, the ill-posedness of the problem poses several challenges to applied researchers as it renders the NPIV estimation problem much more difficult compared to the standard nonparametric regression.

From the practitioner's standpoint, the ill-posed inverse problem limits what can be learnt about γ_0 leading to noisy estimates. The level of ill-posedness is associated with the amount of information the data contain about the structural function and how accurately it can be estimated. Horowitz (2011) points out that only low-order approximation terms can be estimated with desirable precision, which is not a fallacy of the estimation method, but rather a characteristic of the estimation problem itself. In other words, there might not be enough variation in instruments to explain the variation in higher-order approximation terms, meaning that we cannot uncover important nonlinearities from the data. Using a simple Gaussian example, Newey (2013) illustrates the connection between the ill-posedness of the problem and the instrument strength. He demonstrates that the stronger the instrument (the higher the reduced form R^2), the lower the variance of estimates of coefficients of higher-order terms relative to coefficients of lower-order terms. As a result, not only regularization is essential to avoid highly variable estimates, especially when the sample size is relatively small, but also is the strength of constructed instruments.

Another implementation concern is the curse of dimensionality which affects all nonparametric estimators. In the NPIV context, this problem becomes more acute due to the ill-posedness and its effect on convergence rates. For example, in the severely ill-posed case, it might not be possible to obtain a polynomial in n rate, only polynomial in $\log(n)$ (see Blundell et al., 2007; Darolles et al., 2011; Chen and Pouzo, 2012). Consequently, even if the estimation problem is not (moderately) high-dimensional, variance of NPIV estimators can be much higher than that of standard nonparametric regression estimators.

2.2.3. *Review of MLIV estimators*

One promising solution to the aforementioned practical concerns is to appeal to the ML literature which offers a plethora of contemporary data-driven algorithms with various regularization schemes. However, standard ML estimators, such as Lasso, boosting, or Neural Networks are unable to pick up causal effects from endogenous regressors (see e.g., Hartford et al., 2017*b*). This is not a surprise, since the goal of ML estimators is to fit the conditional expectation $\mathbb{E}[Y|X]$, rather than to estimate the structural function γ or any causal effects associated with its shape. We provide an example illustrating this phenomenon in Appendix B.1.

However, despite standard ML algorithms fail in presence of endogeneity, there is a new line of research in machine learning and computer science communities that offers a series of new algorithms that both address endogeneity and can be applied in high-dimensional environments. These MLIV algorithms are data-driven and exploit sophisticated regularization schemes that allow to solve the ill-posed problem while maintaining functional form flexibility.

MLIV estimators can be split into three categories: (i) primal, (ii) dual, and (iii) minimax methods. Primal methods build upon the standard primal formulation of the NPIV estimation problem. It means that in population γ_0 solves

$$\gamma_0 = \underset{\gamma \in \Gamma}{\operatorname{argmin}}\, \mathbb{E}[(Y - \mathbb{E}[\gamma(X)|Z])^2]. \tag{2.2}$$

This is the exact problem the series NPIV estimator solves as well. Hartford et al. (2017*b*) were the first one to suggest using ML to estimate γ in the NPIV setting. Instead of modeling the first stage, they use a Neural Network to model the conditional distribution of endogenous regressors given instruments. Then they plug the estimated pdf in the sample analog of (2.2) and fit another Neural Network to estimate γ. The Double Lasso estimator of Gold et al. (2020) can be seen as a nonparametric series estimator with Lasso in both first and second stages. The Kernel IV (KIV) regression of Singh et al. (2019) is a very powerful estima-

tor that allows to easily deal with high-dimensional inputs without explicitly constructing basis functions or features, which is achieved using the kernel trick. The estimation procedure can be seen a nonlinear generalization of the standard 2SLS estimator, where in both stages instead of the linear regression we run the regularized kernel regression. Bakhitov and Singh (2021) propose a boosting based algorithm to estimate the structural function. The algorithm is very intuitive and resembles an iterative version of the standard 2SLS estimator. Moreover, the approach is data driven, meaning that the researcher does not have to make a stance on neither the form of the target function approximation nor the choice of instruments.

The second group of algorithms focuses on the dual formulation of the estimation problem[12]. The Dual IV (Muandet et al., 2019) uses the dual form of the NPIV estimation problem in (2.2). There are several advantages to using the dual formulation as it collapses the two-stage estimation problem to a one-stage problem. It means, first, that the target function is identified under weaker conditions, completeness is no longer needed, and second, there is no need to model the conditional distribution of X given Z. Bennett et al. (2019) consider the dual version of the GMM IV problem, which can be though of as a natural extension of the Dual IV framework.

Finally, algorithms in the last group are based off of the minimax approach of Dikkala et al. (2020). The main idea is to use violations of the unconditional moment condition as the criterion function, i.e.

$$\gamma_0 = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \max_{f \in \mathcal{F}} \mathbb{E}[(Y - \gamma(X)) f(Z)]. \tag{2.3}$$

Note that the minimax problem in (2.3) does not involve the conditional expectation similar to the dual formulation. Combined with various penalties the minimax criterion function gives rise to a plethora of algorithms to estimate γ. Despite having a different criterion function, the minimax estimator can be asymptotically interpreted as the minimum distance sieve estimator of Chen and Pouzo (2012). However, the formulation is more general and does not restrict Γ and $\mathcal{F}$ to be linear sieve spaces.

In practice, however, the structural function itself is rarely an object of interest, rather it is some economically meaningful object like average partial effects. Consider, for example, a demand estimation problem. The demand level itself does not bear a lot of economic meaning while objects like partial effects of demand shifters, consumer surplus, price and income elasticities or diversion ratios are potential objects of interest. These quantities are functionals of the structural function.

[12]We do not present the dual formulation here as it involves additional derivations. We refer the reader to Muandet et al. (2019) for more details.

2.3. Learning functionals of MLIV estimators

2.3.1. Functionals of interest and economic examples

This paper focuses on estimation and inference on functionals of a flexible (i.e. nonparametric) structural function γ_0 in presence of endogenous regressors, i.e. within the framework of the nonparametric instrumental variables model. Let $W_i \equiv (Y_i, X_i, Z_i)$ be a data observation. Let $m(W, \gamma)$ denote a functional of γ that depends on an observation W. We consider parameters of interest of the form

$$\theta_0 = \mathbb{E}[m(W, \gamma_0)].$$

For expositional convenience, in this Section we will focus on functionals that depend linearly on γ. In Section 2.6 we extend our results to nonlinear functionals. The object of interest θ_0 is an expectation of some functional $m(W, \gamma_0)$ over the data distribution. Hence, we are interested in mean effects, which restricts a set of possible functionals of interest, such as, for example, a simple evaluation functional $\theta_0 = \gamma_0(\bar{X})$, where $\bar{X} \in \text{supp}(X)$. However, our framework is still general enough and covers a wide range of economically important objects.

Below, we give several examples of the types of objects under consideration, including both linear and nonlinear functionals.

Example 1. Weighted average derivative.
In this example, X is a vector of continuous endogenous regressors and

$$\theta_0 = \mathbb{E}\left[\omega(X)\frac{\partial \gamma_0(X)}{\partial X_1}\right],$$

which is a weighted average derivative of γ_0 with respect to X_1 with known weight $\omega(X)$ as in Ai and Chen (2007). Here $m(W, \gamma) = \omega(X)\partial\gamma(X)/\partial X_1$, which is linear in γ. When $\omega(X) = 1$, θ_0 becomes an average partial effect of X_1 on $\gamma_0(X)$.

Example 2. Average policy effect.
The object of interest here is the average effect of changing the covariates according to some transformation $x \mapsto g(x)$,

$$\theta_0 = \mathbb{E}[\gamma_0(g(X)) - \gamma_0(X)],$$

where $m(W, \gamma) = \gamma_0(g(X)) - \gamma_0(X)$ is a linear functional. Thus, θ_0 measures the average policy effect of a counterfactual change of covariate values.

Example 3. Average consumer surplus (CS) and deadweight loss (DWL).
This example is based on Hausman and Newey (1995) and its adaptation to the NPIV setting by Chen and Christensen (2018). Here, $X = (P, I, X_2)$, where P is product price, which is potentially endogenous, I is consumer income, and X_2 includes additional covariates. Let $S(p^0, \iota, x_2)$ denote the exact CS from a price change from p^0 to p^1 at income level ι and covariate values x_2. Then $S(p^0, \iota, x_2)$ is a solution to

$$\frac{\partial S(p(u), \iota, x_2)}{\partial u} = -\gamma_0(p(u), \iota - S(p(u), \iota, x_2), x_2)\frac{\partial p(u)}{\partial u}, \quad S(p(1), \iota, x_2) = 0,$$

where $p : [0, 1] \mapsto \mathbb{R}$ is a twice continuously differentiable price path with $p(0) = p^0$ and $p(1) = p^1$. Let $D(p^0, \iota, x_2)$ denote the corresponding DWL functional given by

$$D(p^0, \iota, x_2) = S(p^0, \iota, x_2) - \left(p^1 - p^0\right)\gamma_0(p^1, \iota, x_2).$$

The objects of interest are

$$\begin{aligned}
\theta_0^{CS} &= \mathbb{E}[\omega(I, X_2)S(p(u), I, X_2)], \\
\theta_0^{DWL} &= \mathbb{E}[\omega(I, X_2)D(p(u), I, X_2)] = \theta_0^{CS} - \mathbb{E}[\omega(I, X_2)(p^1 - p^0)\gamma(p^1, I, X_2)],
\end{aligned}$$

where ω is a weighting function that does not depend on the price level. Unless demand is independent of income, the exact CS and DWL are typically nonlinear functionals of γ_0.

2.3.2. Orthogonal moment condition

Suppose that we are given $\hat{\gamma}$, an MLIV estimator of γ_0. A natural approach to estimate θ_0 is to simply plug-in $\hat{\gamma}$ into m and replace the expectation with the sample average,

$$\hat{\theta}^{\text{plug-in}} = \frac{1}{n}\sum_{i=1}^{n} m(W, \hat{\gamma}).$$

However, the plug-in estimator will not be root-n consistent if the first-order bias does not vanish at root-n rate, which is the case when $\hat{\gamma}$ involves regularization and/or model selection (Chernozhukov, Escanciano, Ichimura, Newey and Robins, 2020). In the NPIV model, regularization is essential to dealing with ill-posedness rendering all NPIV/MLIV estimators regularized estimators.

Figure 2 illustrates the issue. The yellow histogram represents the simulated distribution of the standardized plug-in estimator, $(\hat{\theta}^{\text{plug-in}} - \theta_0)/\text{std}(\hat{\theta}^{\text{plug-in}})$. The estimator is badly bias, shifted to much to the right relative to zero. Moreover, the shape of the distribution is quite different from the standard normal distribution (depicted by the red curve), which would

approximate the asymptotic distribution if bias was negligible. In contrast, the simulated distribution of the standardized debiased estimator that we propose illustrates that the estimator is approximately unbiased (centered around zero) and well approximated by the standard normal distribution, which insures the validity of the inference procedure.

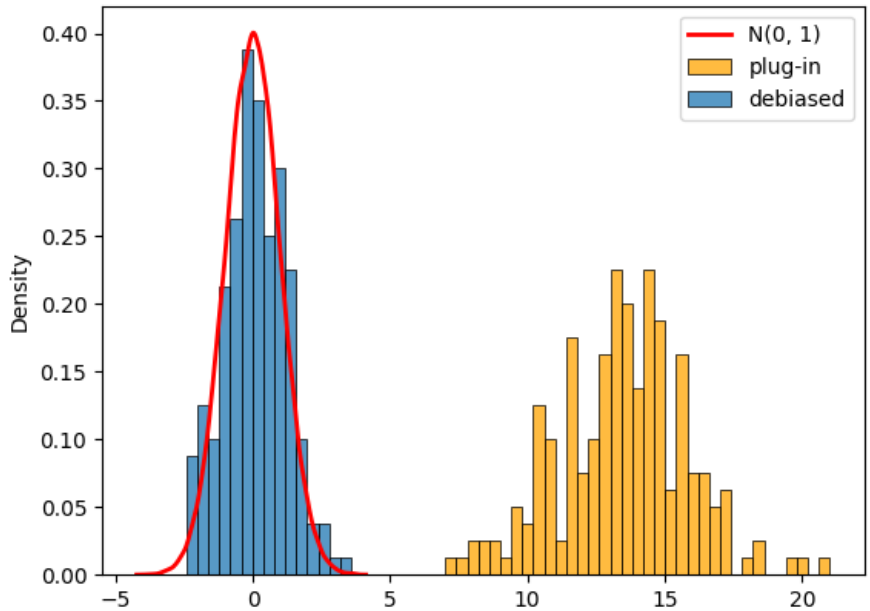

Figure 2: Distributions of plug-in and debiased estimates.

The reason for the plug-in estimator to be affected by the first-order bias is the fact that the moment condition defining θ_0 is not orthogonal to local perturbations of γ around γ_0. Namely, let δ be a local perturbation around γ_0, then the Gateaux derivative in the direction δ is

$$\frac{\partial}{\partial \tau}\mathbb{E}[m(W, \gamma_0 + \tau\delta)]\bigg|_{\tau=0} = \mathbb{E}[m(W, \delta)] \neq 0.$$

Thus, obtaining an orthogonal moment condition is a crucial step for establishing our results.

We consider functionals $m(W, \gamma)$ where there exists a function $\alpha_0(Z)$ with $\mathbb{E}[\alpha_0^2(Z)] < \infty$ and

$$\mathbb{E}[m(W, \gamma)] = \mathbb{E}[\alpha_0(Z)\gamma(X)] \text{ for all } \gamma \text{ with } \mathbb{E}[\gamma^2(X)] < \infty. \tag{2.4}$$

As discussed in Ichimura and Newey (2017), if there exists $v(X)$ with $\mathbb{E}[v^2(X)] < \infty$ and $\mathbb{E}[m(w, \gamma)] = \mathbb{E}[v(X)\gamma(X)]$, then the existence of $\alpha_0(Z)$ requires $v(X) = \mathbb{E}[\alpha_0(Z)|X]$. As

pointed out in Severini and Tripathi (2012), this is a necessary condition for root-n estimability of θ_0. Moreover, by the Riesz representation theorem, the existence of such $\alpha_0(Z)$ is equivalent to $\mathbb{E}[m(W,\gamma)]$ being a mean square continuous functional of γ. Henceforth, we refer to $\alpha_0(Z)$ as a Riesz representer. Newey (1994) shows that mean square continuity of $\mathbb{E}[m(W,\gamma)]$ is equivalent to the semiparametric efficiency bound of θ_0 being finite. Thus, our approach focuses on regular functionals. Similar uses of the Riesz representation theorem can be found in Ai and Chen (2007), Ackerberg et al. (2014), Hirshberg and Wager (2020), and CNS among others.

Ichimura and Newey (2017) establish the form of the orthogonal moment function for NPIV estimators

$$\psi(W,\theta,\gamma,\alpha) = m(W,\gamma) - \theta + \alpha(Z)[Y - \gamma(X)], \tag{2.5}$$

where $\alpha(Z)[Y - \gamma(X)]$ is the influence function. Note that the moment function in (2.5) is Neyman orthogonal to local perturbations (δ,β) of (γ_0,α_0) such that

$$\frac{\partial}{\partial\tau}\mathbb{E}[\psi(W,\theta,\gamma_0+\tau\delta,\alpha_0+\tau\beta)]\bigg|_{\tau=0} = \mathbb{E}[m(W,\delta)] - \mathbb{E}[\alpha_0(Z)\delta(X)] + \mathbb{E}[(Y-\gamma_0(X))\beta(Z)] = 0,$$

where the first two terms cancel out by the Riesz representation theorem and the last term is zero by the exogeneity condition. This property makes the orthogonal moment condition an excellent basis for constructing a debiased estimator of θ_0 in the NPIV setting where estimators are typically regularized. Similar uses of the Neyman-orthogonal moment condition can be found in Chen et al. (2021) for NPIV sieve estimators and in Gautier and Rose (2021) for the high-dimensional linear IV regression.

Moreover, the exogeneity condition and iterated expectations imply

$$\mathbb{E}[\alpha(Z)(Y - \gamma_0(X))] = \mathbb{E}[\alpha(Z)\mathbb{E}[Y - \gamma_0(X)|Z]] = 0$$

for any $\alpha(Z)$, meaning that the expectation of the influence function is zero regardless of α. This implies

$$\mathbb{E}[\psi(W,\theta_0,\gamma_0,\alpha)] = \mathbb{E}[m(W,\gamma_0)] - \theta_0 + \mathbb{E}\left[\alpha(Z)[Y - \gamma_0(X)]\right] = 0,$$

which allows us to use (2.5) to estimate θ_0. The debiased estimator $\hat{\theta}$ can be constructed by plugging in $\hat{\gamma}$ and $\hat{\alpha}$ into the moment function $\psi(W,\theta,\gamma,\alpha)$ in place of γ and α and solving for $\hat{\theta}$ from setting the sample moment $\psi(W,\theta,\hat{\gamma},\hat{\alpha})$ to zero.

Note that the debiased estimator $\hat{\theta}$ requires an estimator of α_0. Typically in the NPIV setting, the form of α_0 is very complicated to derive or even unknown. Consider the weighted

average derivative example from above. The RR is a solution to the following integral equation

$$\mathbb{E}[\alpha_0(Z)|X] = -\frac{\partial\{f_0(X)\omega(X)\}/\partial X_1}{f_0(X)},$$

where $f_0(X)$ is the marginal pdf of X. As a result, it is desirable to have a flexible approach for automatic estimation of the RR. The next subsection describes how to construct such an estimator.

2.3.3. Estimation of the Riesz representer

Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) show that we can exploit the orthogonality of the debiased moment function $\psi(W, \theta, \gamma, \alpha)$ to estimate α_0. The Gateaux derivative of $\psi(W, \theta, \gamma, \alpha)$ in the direction δ is

$$\begin{aligned}
\mathbb{E}[\psi_\gamma(W, \theta_0, \delta, \alpha_0)] &= \frac{\partial}{\partial\tau}\mathbb{E}[\psi(W_i, \theta_0, \gamma_0 + \tau\delta, \alpha_0)]\Big|_{\tau=0} \\
&= \frac{\partial}{\partial\tau}\mathbb{E}\left[m(W, \gamma_0 + \tau\delta) - \theta_0 + \alpha_0(Z)[Y - \gamma_0(X) - \tau\delta(X)]\right]\Big|_{\tau=0} \\
&= \mathbb{E}[m(W, \delta) - \alpha_0(Z)\delta(X)] = 0, \qquad (2.6)
\end{aligned}$$

where the last equality comes from $m(W,\gamma)$ being linear in γ. This can be thought of as a population moment condition for α_0.

Several recent papers propose different Riesz representer estimators based on the moment condition in (2.6) under exogeneity. CNS use minimum distance Lasso and Dantzig estimators. A recent follow-up paper by Chernozhukov et al. (2021) extend the CNS' approach and use a neural network to estimate α_0. Chernozhukov, Newey, Singh and Syrgkanis (2020) take a different approach and allow for a more general learner of α_0 based on the minimax framework of Dikkala et al. (2020). It is important to highlight once more that the aforementioned approaches allow for estimation of the Riesz representer only under exogeneity, when α_0 is a function of X rather than Z.

We assume that the Riesz representer estimator takes the form $\hat{\alpha} = b(Z)'\hat{\rho}$, where $b(Z)$ is a p-dimensional dictionary of basis functions with p being possibly much larger than n. Let $d(X)$ be a q-dimensional dictionary of basis functions that represent deviations from γ_0. Using $d(X)$, we can construct a vector of moment conditions to estimate ρ. Let $d_j(X)$ be an element of $d(X)$, then we can form a sample moment condition corresponding to the population moment condition (2.6) by replacing the expectation with a sample average and

$\alpha_0(Z)$ with $b(Z)'\rho$ to obtain

$$\hat{\psi}_\gamma(d_j, \rho) = \frac{1}{n}\sum_{i=1}^{n}\left\{m(W_i, d_j) - d_j(X_i)b(Z_i)'\rho\right\} = 0, \quad j = 1, \ldots, q. \tag{2.7}$$

Note that we require $q \geq p$ to ensure identification and estimability of ρ.

To allow for a high-dimensional α specification, we follow Caner and Kock (2018) and use the penalized GMM (PGMM) framework. Let $\hat{\psi}_\gamma(\rho) = (\hat{\psi}_\gamma(d_1, \rho), \ldots, \hat{\psi}_\gamma(d_q, \rho))'$ where $\hat{\psi}_\gamma(d_j, \rho)$ is defined in (2.7). Then a solution to the PGMM problem takes the form

$$\hat{\rho}_L = \underset{\rho \in \mathbb{R}^p}{\operatorname{argmin}}\, \hat{\psi}_\gamma(\rho)'\hat{\Omega}_q\hat{\psi}_\gamma(\rho) + 2\lambda_n|\rho|_1, \tag{2.8}$$

where $\hat{\Omega}_q = \hat{\Omega}/q$, $\hat{\Omega}$ is a $q \times q$ positive semi-definite matrix, and $2\lambda_n|\rho|_1$ is a penalty term. This framework allows for $q \geq p > n$, and basically is a Lasso extension of the standard GMM.

Let $\hat{G} = \frac{1}{n}\sum_{i=1}^{n} d(X_i)b'(Z_i)$ and $\hat{M} = \frac{1}{n}\sum_{i=1}^{n} m(W_i, d)$ be unbiased estimators of $G = \mathbb{E}[d(X)b'(Z)]$ and $M = \mathbb{E}[m(W, d)]$, respectively. Then we can rewrite (2.8) in matrix form as

$$\hat{\rho}_L = \underset{\rho \in \mathbb{R}^p}{\operatorname{argmin}}(\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda_n|\rho|_1. \tag{2.9}$$

The estimator $\hat{\rho}_L$ can be interpreted as a minimum distance version of the high-dimensional GMM estimator of Caner and Kock (2018). Note that we cannot use the standard optimal weight matrix as for the low-dimensional GMM due to its rank deficiency. Implementation details can be found in Appendix B.3.

2.3.4. Informal preview of estimation and inference results

The estimation procedure can be summarized in the following pseudo-algorithm:

1. We follow CNS and use cross-fitting to avoid (i) potentially severe finite sample bias due to the double use of data and (ii) regularity conditions based on $\hat{\gamma}$ and $\hat{\alpha}$ being in Donsker class, which ML estimators are usually not. Assuming the data $\{W\}_{i=1}^n$ is $i.i.d.$, let I_ℓ, $\ell = 1, \ldots, L$, be a partition of the observation index set $\{1, \ldots, n\}$ into L distinct subsets of about equal size. Let n_ℓ denote the number of observations in fold ℓ.

2. For each data fold $\ell = 1, \ldots, L$, we obtain estimates $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ that are constructed from the observations not in I_ℓ. In particular, the RR estimate is of the form $\hat{\alpha}_\ell =$

$b(Z)'\hat{\rho}_\ell$, where

$$\hat{\rho}_\ell = \underset{\rho \in \mathbb{R}^p}{\text{argmin}}(\hat{M}_\ell - \hat{G}_\ell \rho)'\hat{\Omega}_q(\hat{M}_\ell - \hat{G}_\ell \rho) + 2\lambda_n |\rho|_1,$$

with $\hat{G}_\ell = \frac{1}{n-n_\ell}\sum_{i \notin I_\ell} d(X_i)b'(Z_i)$ and $\hat{M}_\ell = \frac{1}{n-n_\ell}\sum_{i \notin I_\ell} m(W_i, d)$.

3. We construct the estimator $\hat{\theta}$ by setting the sample average of $\psi(W, \theta, \hat{h}_\ell, \hat{\alpha}_\ell)$ to zero and solving for θ. This estimator $\hat{\theta}$ and the associated asymptotic variance estimator $\hat{V}$ have the following explicit forms

$$\hat{\theta} = \frac{1}{n}\sum_{\ell}^{L}\sum_{i \in I_\ell}\{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(Z_i)[Y_i - \hat{\gamma}_\ell(X_i)]\} \tag{2.10}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell}\hat{\psi}_{i\ell}^2,\ \hat{\psi}_{i\ell} = m(W_i, \hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(Z_i)[Y_i - \hat{\gamma}_\ell(X_i)].$$

Next, we informally discuss the key conditions behind the asymptotic normality result. Since $\hat{\theta}$ is constructed by plugging-in $\hat{\gamma}$ and $\hat{\alpha}$ in the orthogonal moment condition, asymptotic properties of $\hat{\theta}$ depend on the asymptotic behavior of $\hat{\gamma}$ and $\hat{\alpha}$. First, to allow for a wide range of MLIV estimators, we assume that $\hat{\gamma}$ satisfies some projected mean square convergence rate condition as an estimator of γ_0. Specifically, we require

$$||T(\hat{\gamma} - \gamma_0)|| = O_p(\kappa_n^\gamma),$$

where κ_n^γ can be slower than root-n rate[13]. As pointed out in Section 2.2.1, it is possible to obtain a fast rate under the projected mean square norm. Hence, it is a weak high-level assumption that can be satisfied by a variety of MLIV estimators such as Double Lasso (Gold et al., 2020), Kernel IV (Singh et al., 2019) and a series of estimators constructed using the minimax framework of Dikkala et al. (2020).

The second condition is the mean square convergence rate of $\hat{\alpha}$. For the ease of exposition, assume that $\hat{\alpha}$ satisfies the following mean square convergence rate condition,

$$||\hat{\alpha} - \alpha_0|| = O_p(\kappa_n^\alpha).$$

We derive an exact expression for κ_n^α in Section 2.4.

Finally, under quite standard regularity conditions asymptotic normality can be established

[13]The result also holds for the standard mean square rate condition, i.e. $||\hat{\gamma} - \gamma_0|| = O_p(\kappa_n^\gamma)$, however, for NPIV/MLIV estimators this rate is slower due to ill-posedness.

provided that

$$\sqrt{n}||\hat{\alpha} - \alpha_0|| \; ||T(\hat{\gamma} - \gamma_0)|| \xrightarrow{p} 0,$$

which is satisfied when $\sqrt{n}\kappa_n^{\gamma}\kappa_n^{\alpha} \to 0$. Hence, there is a trade-off between the convergence rates of $\hat{\gamma}$ and $\hat{\alpha}$. It is possible to allow for a slower convergence rate of $\hat{\gamma}$ at the expense of a faster convergence rate of $\hat{\alpha}$ and vice versa.

2.4. Properties of the PGMM estimator

In this section we provide the mean square convergence rate for the PGMM estimator $\hat{\alpha}$ which is necessary for the asymptotic analysis of $\hat{\theta}$. We start by introducing some conditions.

Assumption 5. There exists a sequence of non-random matrices Ω such that

$$||\hat{\Omega} - \Omega||_{\ell_\infty} = o_p(1) \quad and \quad ||\Omega||_{\ell_\infty} \leq C < \infty$$

for some constant C.

The first part of Assumption 5 is pretty standard and requires a consistent estimate of the weight matrix. The second part of the assumption, as discussed in Caner and Kock (2018), might be restrictive as it requires a high-dimensional matrix to be uniformly bounded in ℓ_∞- norm, but for the notational convenience we keep it. The analysis in the paper will still go through if we switch to a diagonal weight matrix as Caner and Kock (2018) suggest.

Note that the convergence rate of the PGMM estimator defined in (2.9) depends on the convergence rates of $\hat{\Omega}$, $\hat{G}$, and $\hat{M}$. Assumption 5 ensures that $\hat{\Omega}$ is consistent. To obtain a convergence rate for $\hat{G}$, we impose the following condition.

Assumption 6. There are constants C_b and C_d such that with probability approaching one,

$$\max_{1\leq j\leq p} |b_j(Z)| \leq C_b \text{ and } \max_{1\leq j\leq q} |d_j(X)| \leq C_d.$$

This condition implies

$$||\hat{G} - G||_\infty = O_p(\varepsilon_n^G), \text{ where } \varepsilon_n^G = \sqrt{\frac{\log(q)}{n}}.$$

Unlike the standard Lasso, the second moment matrix convergence rate depends on the number of moments, i.e. the number of elements in $d(X)$, rather than the number of elements in $b(Z)$.

Let us hypothesize a convergence rate for $\hat{M}$.

Assumption 7. There is ε_n^M such that

$$||\hat{M} - M||_\infty = O_p(\varepsilon_n^M),\ \varepsilon_n^M \to 0.$$

Next, we proceed by following CNS and impose a sparse approximation condition for α_0.

Assumption 8. There exist $C > 1$ and $\bar{\rho}$ with $\bar{s}$ non-zero elements such that

$$||\alpha_0 - b'\bar{\rho}||^2 \leq C\bar{s}\varepsilon_n^2,$$

where $\varepsilon_n = \max\{\varepsilon_n^G, \varepsilon_n^M\}$.

Intuitively, this assumption controls the squared approximation error from using the linear combination $b'\bar{\rho}$ to approximate α_0. Note that Assumption 8 does not necessarily require α_0 to be equal to the linear combination of $\bar{s}$ terms, it states that there exists a sparse $\bar{\rho}$ with $\bar{s}$ non-zero elements such that the approximation error is bounded by $C\bar{s}\varepsilon_n^2$. In other words, Assumption 8 is general enough to accommodate both exact and approximate sparsity of α_0. Approximate sparsity allows for a large number of potential regressors (possibly much larger than the sample size) when relatively few important regressors give a good approximation but the identity of those few is not known, which is different from a standard series approximation where typically the first $\bar{s}$ regressors are assumed to achieve a good approximation (Bradic et al., 2021). Thus, very sparse approximations allow to keep $\bar{s}$ relatively small which results in faster convergence rates. For a more detailed discussion of approximation bias conditions we refer the reader to CNS.

Let $S = \{1, \ldots, p\}$, S_ρ be a subset of S with $\rho_j \neq 0$, and S_ρ^c be the complement of S_ρ in S. Let ρ_L be the population coefficients, i.e.

$$\rho_L = \underset{\rho \in \mathbb{R}^p}{\operatorname{argmin}} (M - G\rho)'\Omega_q(M - G\rho) + 2\varepsilon_n|\rho|_1.$$

The PGMM estimator $\hat{\rho}_L$ estimates the population coefficients ρ_L, which in turn might be different from the approximation coefficients $\bar{\rho}$. The following condition is essential to derive the oracle inequality for $\hat{\rho}_L$, and hence, the convergence rate for $\hat{\alpha}_L = b'\hat{\rho}_L$.

Assumption 9. Let $G'\Omega_q G$ have its largest eigenvalue uniformly bounded in n and

$$\phi^2(s) = \inf\left\{\frac{\delta' G'\Omega_q G\delta}{||\delta_{S_\rho}||^2} : \delta \in \mathbb{R}^p \backslash \{0\},\ |\delta_{S_\rho^c}|_1 \leq 3|\delta_{S_\rho}|_1, |S_\rho| \leq s\right\} > 0.$$

Assumption 9 is the modified population restricted eigenvalue condition as in Caner and Kock (2018). To accommodate for the PGMM estimator the condition is imposed on $G'\Omega_q G$ rather than $\mathbb{E}[b(Z)b'(Z)]$ as in the classic restricted eigenvalue condition of Bickel et al. (2009). Showing that its empirical counterpart is bounded uniformly away from zero will be used to put a bound on the estimation error of $\hat{\alpha}_L$.

Assumption 10. There is $C > 0$ such that with probability approaching one,

$$\max_{1 \leq j \leq q} |m(W, d_j)| \leq C.$$

This condition is needed to put a bound on $||M||_\infty$ which is necessary to establish the oracle inequality for $\hat{\rho}_L$, and hence, the convergence rate for $\hat{\alpha}_L$. Moreover, note that by Assumption 10, $\varepsilon_n = \varepsilon_n^M = \varepsilon_n^G = \sqrt{\log(q)/n}$. This simplifies the analysis, but is not necessary for establishing the results below. Also, let $|\bar{\rho}|_1 \leq \bar{A} < \infty$. We can allow for the norm to grow with n at a certain rate, however, it does not change the main results, hence, for simplicity we put a bound on $|\bar{\rho}_1|$.

Theorem 4. If Assumptions 6–10 are satisfied and $\varepsilon_n = o(\lambda_n)$, then

$$||\hat{\alpha}_L - \alpha_0||^2 = O_p(\kappa_n^\alpha) \text{ where } \kappa_n^\alpha = \bar{s}^2 \lambda_n^2.$$

The presence of endogeneity results in a slower rate of convergence for the RR estimator compared to the exogenous counterpart in CNS. The MD Lasso estimator of CNS converges at $\bar{s}\lambda_n^2$ rate, while the PGMM estimator is slower by a factor of $\bar{s}$. Note that the convergence rate only depends on the number of approximation elements $\bar{s}$, but is independent of the number of relevant moments.

Example 4. Consider the approximately sparse case where there are constants C and $\xi > 0$ such that

$$||\alpha_0 - b'\bar{\rho}||^2 \leq C(\bar{s})^{-\xi}.$$

Let $\bar{s} \leq C\varepsilon_n^{-2/(1+2\xi)}$. Then Assumption 8 is satisfied with

$$||\alpha_0 - b'\bar{\rho}||^2 = O\left(\varepsilon_n^{2\xi/(1+2\xi)}\right)$$

and

$$||\hat{\alpha}_L - \alpha_0||^2 = O_p\left(\varepsilon_n^{-4/(1+2\xi)}\lambda_n^2\right).$$

Suppose that $\varepsilon_n = \sqrt{\log(q)/n}$ and let a_n be a sequence converging to infinity very slowly

with n, e.g. $a_n = \log(\log(n))$. Then for $\lambda_n = a_n \varepsilon_n$,

$$\varepsilon_n^{-4/(1+2\xi)} \lambda_n^2 = \left(\frac{\log(q)}{n}\right)^{-\frac{2}{1+2\xi}+1} a_n^2 = \left(\frac{\log(q)}{n}\right)^{\frac{2\xi-1}{1+2\xi}} a_n^2.$$

This rate is slower than the CNS rate

$$\left(\frac{\log(p)}{n}\right)^{\frac{2\xi}{1+2\xi}} a_n^2.$$

However, this difference becomes negligible for large enough ξ.

2.5. Asymptotic properties of linear functionals

In this Section, we provide conditions ensuring root-n consistency and asymptotic normality of the debiased estimator $\hat{\theta}$. Under the specified conditions, we can do inference in a standard way. First, we focus on linear functionals and then provide additional conditions to extend the results to nonlinear functionals in Section 2.6.

We impose the following conditions.

Assumption 11. $\alpha_0(z)$ and $\mathbb{E}\left[[y - \gamma_0(x)]^2 | z\right]$ are bounded and $\mathbb{E}\left[m(w, \gamma_0)^2\right] < \infty$.

This assumption is purely technical, and we maintain it for simplicity.

Assumption 12. $\int [m(w, \hat{\gamma}) - m(w, \gamma_0)]^2 F_0(dw) \xrightarrow{p} 0$ and $||\hat{\gamma} - \gamma_0|| \xrightarrow{p} 0$.

Assumption 13. $||T(\hat{\gamma} - \gamma_0)|| = O_p(\kappa_n^\gamma)$ with $\kappa_n^\gamma \to 0$.

Assumption 12 allows for estimators $\hat{\gamma}$ that are mean square consistent. Assumption 13 requires $\hat{\gamma}$ to converge to γ_0 in the projected norm at a rate equal to κ_n^γ which is typically slower than root-n. Note that this condition is weaker than convergence in standard mean square norm (see Section 2.2.1). This specification is general enough and allows for various MLIV estimators.

Assumption 14. $\varepsilon_n = o(\lambda_n)$ and $\sqrt{n}\kappa_n^\alpha \kappa_n^\gamma \to 0$.

This condition is sufficient to guarantee $\sqrt{n}||\hat{\alpha}_L - \alpha_0|| \, ||T(\hat{\gamma} - \gamma_0)|| \xrightarrow{p} 0$, leading to asymptotic normality of $\hat{\theta}$. Recall Example 4, in which case Assumption 14 requires

$$\sqrt{n}\bar{s}\lambda_n \kappa_n^\gamma = O\left(n^{1/2}\left(\sqrt{\frac{\log(q)}{n}}\right)^{\frac{2\xi-1}{1+2\xi}} a_n \kappa_n^\gamma\right) \to 0. \tag{2.11}$$

Suppose $\kappa_n^\gamma = n^{-d_\gamma}$ with $d_\gamma > 0$. Then condition (2.11) implies

$$\frac{2\xi - 1}{2(1+2\xi)} + d_\gamma > \frac{1}{2}.$$

Thus, as in CNS, there is a trade-off between between ξ, which determines how sparse the approximation is, and d_γ, the convergence rate of $\hat{\gamma}$ in the projected norm. Note that it forces $\hat{\gamma}$ to converge faster compared to CNS whose rate condition is $2\xi/(1+2\xi)+d_\gamma > 1/2$, which is a consequence of the lower rate of convergence of $\hat{\alpha}_L$. However, for large enough ξ, d_γ can still be arbitrary small.

Theorem 5. If Assumptions 6–14 are satisfied, then for $\psi_0(w) = m(w, \gamma_0) - \theta_0 + \alpha_0(z)[y - \gamma(x)]$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V) \text{ and } \hat{V} \xrightarrow{p} V = \mathbb{E}\left[\psi_0^2(w)\right].$$

2.6. Nonlinear functionals

The results from Section 2.5 can be extended to allow for estimation of $\theta_0 = \mathbb{E}[m(W, \gamma_0)]$ for nonlinear $m(W, \gamma)$. The estimator is similar to the linear case except we estimate the RR of the linearization of $m(W, \gamma)$ leading to a different $\hat{M}$ needed. In this Section, we show how to construct such an estimator and provide additional conditions that are sufficient for valid asymptotic inference for nonlinear functionals. As we mentioned in the introduction, due to nonlinearity of $m(W, \gamma)$, we have to impose restrictions on the convergence rate of $\hat{\gamma}$ in terms of the standard mean square norm, not the projected norm as in the linear case. We provide more details below.

To account for nonlinearity of $m(W, \gamma)$ in γ, we assume linearity of the Gateaux derivative of a nonlinear functional (see Chernozhukov, Newey and Singh, 2018). To be more precise, let ζ be a deviation from γ. We assume that $m(W, \gamma)$ is Gateaux differentiable with the derivative $D(W, \gamma, \zeta)$, meaning that

$$D(W, \gamma, \zeta) = \frac{d}{d\tau} m(W, \gamma + \tau\zeta)\Big|_{\tau=0}$$

for a scalar τ, and that $D(W, \gamma, \zeta)$ is linear in ζ. Moreover, assume that $\alpha_0(Z)$ satisfies

$$\mathbb{E}[D(W, \gamma_0, \zeta)] = \mathbb{E}[\alpha_0(Z)\zeta(X)], \text{ for all } \zeta(X) \text{ with } \mathbb{E}[\zeta^2(X)] < \infty. \tag{2.12}$$

In other words, Equation (2.12) implies that $D(W, \gamma, \zeta)$ is a mean-square continuous functional of ζ, which corresponds to Assumption 3 of Ichimura and Newey (2017), meaning that $\alpha_0(Z)$ is a Riesz representer of the Gateaux derivative of $m(W,\gamma)$ with respect

to γ evaluated at $\gamma = \gamma_0$. Thus, by the Riesz representation theorem, for $D(W, \gamma_0, d) = (D(W, \gamma_0, d_1), \ldots, D(W, \gamma_0, d_q))'$,

$$M = \mathbb{E}[D(W, \gamma_0, d)] = \mathbb{E}[\alpha_0(Z)d(X)].$$

We can construct an estimator $\hat{\theta}$ exactly like in Equation (2.10) except we need a different estimator of $\alpha_0(Z)$ based on (2.12). Despite γ enters $m(W, \gamma)$ nonlinearly, the estimator will still have zero first-order bias and be root-n consistent and asymptotically normal under suficient regularity conditions. See Newey (1994), Ichimura and Newey (2017), and Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) for more details.

An estimator $\hat{\alpha}_\ell$ can be constructed exactly as described in Section 2.3.3 except being based on a different $\hat{M}_\ell$, where it is convenient to bring back the ℓ subscript. Let $\hat{\gamma}_{\ell,\ell'}$ be based on observations not in either I_ℓ or $I_{\ell'}$, then the unbiased estimator $\hat{M}_\ell$ is given by

$$\begin{aligned}
\hat{M}_\ell &= (\hat{M}_{\ell 1}, \ldots, \hat{M}_{\ell q.})' \\
\hat{M}_{\ell j} &= \frac{1}{n - n_\ell} \sum_{\ell' \neq \ell} \sum_{i \in I_\ell} D(W_i, \hat{\gamma}_{\ell,\ell'}, d_j),
\end{aligned}$$

where $\hat{M}_{\ell j}$ is the Gateaux derivative of the moment function with respect to γ in the direction of the jth dictionary function. This estimator uses further sample splitting where $\hat{M}$ is constructed by averaging over observations that are not used in $\hat{\gamma}_{\ell,\ell'}$. This additional sample splitting allows $\hat{M}_\ell$ to depend on an estimator of γ as required when $m(W, \gamma)$ is nonlinear in γ.

To establish the convergence rate for $\hat{M}_\ell$, we impose the following condition.

Assumption 15. There exist $C, \varepsilon > 0$ such that for any γ with $||\gamma - \gamma_0|| \leq \varepsilon$:

(i) $\max_{1 \leq j \leq q} |D(W, \gamma, d_j)| \leq C$;

(ii) $\sup_{1 \leq j \leq q} |\mathbb{E}[D(W, \gamma, d_j) - D(W, \gamma_0, d_j)]| \leq C||\gamma - \gamma_0||$.

Lemma 4. Suppose that $||\hat{\gamma}_{\ell,\ell'} - \gamma_0|| = O_p(\kappa_n^\gamma)$ for $\ell, \ell' = 1, \ldots, L$, and Assumption 15 is satisfied, then

$$||\hat{M}_\ell - M_\ell||_\infty = O_p(\kappa_n^\gamma).$$

As CNS point out, the presence of the initial estimator $\hat{\gamma}_{\ell,\ell'}$ in $\hat{M}_\ell$ makes the convergence rate of $||\hat{M}_\ell - M_\ell||_\infty$ slower, κ_n^γ instead of $\sqrt{\log(q)/n}$. Thus, $\varepsilon_n = \varepsilon_n^M = \kappa_n^\gamma$, which requires λ_n to converge to zero slightly slower than κ_n^γ. This also affects the convergence rate condition

in Assumption 14. Let us illustrate this effect using the set-up from Example 4. Under $\kappa_n^\gamma = n^{-d_\gamma}$ and $\varepsilon_n = n^{-d_\gamma}$, Assumption 14 requires

$$n^{1/2}\bar{s}\lambda_n n^{-d_\gamma} = O\left(n^{1/2} n^{-d_\gamma \frac{4\xi}{1+2\xi}}\right) \to 0,$$

implying

$$d_\gamma \frac{4\xi}{1+2\xi} > \frac{1}{2}.$$

This condition will be satisfied for any $d_\gamma > 1/4$, given ξ is large enough. The result is similar to CNS whose rate condition is $d_\gamma(4\xi+1)/(1+2\xi) > 1/2$. When $\kappa_n^\gamma = \log(n)^{-d_\gamma}$, it is required that

$$n^{1/2}\log(n)^{-d_\gamma \frac{4\xi}{1+2\xi}} \to 0.$$

For large enough ξ, it implies that d_γ must satisfy $\log(n)^{-d_\gamma} = o(n^{-1/4})$.

Assumption 16. There exist $C, \varepsilon > 0$ such that for any γ with $||\gamma - \gamma_0|| \leq \varepsilon$,

$$|\mathbb{E}[m(W,\gamma) - m(W,\gamma_0) - D(W,\gamma_0,\gamma-\gamma_0)]| \leq C||\gamma-\gamma_0||^2.$$

This condition controls the size of the linearization remainder in a linearization using the Gateaux derivative. It implies that $\mathbb{E}[m(W,\gamma)]$ is Frechet differentiable in $||\gamma-\gamma_0||$ at γ_0 with derivative $\mathbb{E}[D(W,\gamma_0\,\gamma-\gamma_0)]$.

Assumption 17. $||\hat{\gamma}-\gamma_0|| = O_p(\kappa_n^\gamma)$ and $n^{1/4}||\hat{\gamma}-\gamma_0|| \xrightarrow{p} 0$.

It is a standard assumption to accommodate for nonlinearity of $m(W,\gamma)$. This might be a very tight restriction to satisfy given overall slow convergence rates of NPIV estimators, especially in the severely ill-posed case. However, as discussed in CNS, it is not known whether it is possible to weaken the $n^{-1/4}$ condition for nonlinear functionals, which goes back to Newey (1994).

Theorem 6. If Assumptions 6–12, 14, and 15–17 are satisfied, then for $\psi_0(w) = m(w,\gamma_0) - \theta_0 + \alpha_0(z)[y-\gamma(x)]$,

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,V), \quad \hat{V} \xrightarrow{p} V = \mathbb{E}\left[\psi_0^2(w)\right].$$

2.7. Monte Carlo

In this Section, we present a simple Monte Carlo exercise illustrating the final sample performance of the approach. We compare the performance of the debiased estimator to the

plug-in estimator.

Our design bases off of the MC design of Newey and Powell (2003), Santos (2012) and Chen and Pouzo (2015), which we modify to allow for multiple regressors and instruments. To be specific, we generate $i.i.d.$ draws

$$\begin{pmatrix} X_{ij} \\ Z_{ij} \\ u_{ij} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right), \quad j = 1, \ldots, k$$

The true structural function is given by

$$\gamma(X_i) = \exp\{-0.5 X_i' X_i\},$$

which is the pdf of a product of k standard normal random variables. The response variable is generated as

$$Y_i = \gamma(X_i) + v_i, \quad v_i = \sum_{j=1}^{k} u_{ij},$$

where v_i is a composite error term. Note that this form of the composite error term implies that the degree of endogeneity, i.e. the correlation between each individual regressor X_j and v diminishes with k. As a result, we do not consider k greater than 10. The functional of interest is a weighted average of the form

$$\theta = \mathbb{E}[w(X)\gamma(X)], \quad w(X) = X'X.$$

We construct dictionaries $b(Z)$ and $d(X)$ using cubic polynomials with interaction terms. Since $dim(X_i) = dim(Z_i) = k$, both dictionaries have the same number of basis functions, i.e. $p = q$. To estimate the structural function γ, we use the Double Lasso estimator of Gold et al. (2020). We run 1000 replications for $k = 2$, 5, 10 and $n = 100$, 500, 1000, and 5000. Estimation is carried out using five-fold $(L = 5)$ cross-fitting.

The results are presented in Table 4. The plug-in estimator is labeled PI, while DB stands for the debiased estimator. Bias is the absolute value of bias, SD is the standard deviation, RMSE is the root mean square error, and Cvg denotes the coverage probability of a 95% nominal confidence interval.

In all cases the debiased estimator has a significantly smaller bias than the plug-in estimator. Moreover, the coverage probabilities for the debiased estimator are pretty close to the nominal level except for $k = 2$, $n = 5000$ case. On the other hand, larger bias of the

Table 4: MC results: weighted average derivative.

		PI Bias	DB Bias	PI SD	DB SD	PI RMSE	DB RMSE	PI Cvg	DB Cvg
$k = 2$	$n = 100$	0.106	0.022	0.313	0.253	0.330	0.254	0.51	0.94
	$n = 500$	0.092	0.028	0.080	0.073	0.122	0.078	0.26	0.94
	$n = 1000$	0.068	0.028	0.058	0.050	0.090	0.058	0.21	0.92
	$n = 5000$	0.044	0.028	0.023	0.023	0.050	0.036	0.07	0.77
$k = 5$	$n = 100$	0.107	0.028	0.259	0.249	0.280	0.251	0.69	0.96
	$n = 500$	0.107	0.042	0.096	0.100	0.144	0.109	0.17	0.95
	$n = 1000$	0.103	0.035	0.068	0.072	0.123	0.080	0.07	0.94
	$n = 5000$	0.070	0.020	0.037	0.034	0.079	0.040	0.04	0.90
$k = 10$	$n = 100$	0.043	0.044	0.374	0.352	0.377	0.355	0.77	0.96
	$n = 500$	0.030	0.009	0.144	0.141	0.147	0.141	0.56	0.96
	$n = 1000$	0.027	0.013	0.096	0.096	0.100	0.097	0.33	0.96
	$n = 5000$	0.029	0.013	0.044	0.046	0.053	0.048	0.06	0.94

plug-in estimator results in poor coverage that is far from the nominal level for all cases. Furthermore, for all cases the debiased estimator has a smaller RMSE, which is due to bias reduction. Overall, our results are similar to Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020), which indicates that our procedure is valid and performs well in practice.

2.8. Conclusion

In this paper, we have given an automatic method of debiasing functionals of machine learners under endogeneity. We have shown how to use a PGMM minimum distance estimator to perform debiasing using only the form the object of interest, without knowing the form of the bias correction term. We allow for a wide range of MLIV estimators that satisfy certain convergence rate conditions. We have shown root-n consistency and asymptotic normality and given a consistent asymptotic variance estimator for both linear and nonlinear functionals. For linear functionals we require MLIV estimators to converge fast enough in the projected mean square norm, while for nonlinear functionals we require fast enough convergence in the standard mean square norm, which is a more stringent requirement due to ill-posedness. Relaxing the convergence rate condition for nonlinear functionals as well extending the approach to irregular functionals are promising directions for future research.

CHAPTER 3: Flexible Demand Estimation using Machine Learning

3.1. Introduction

Demand estimation for differentiated products plays a central role in modern empirical industrial organization. The groundbreaking work of Berry (1994) and Berry et al. (1995) (henceforth, BLP) provides an important framework for analyzing aggregate demand by jointly modeling consumer preference heterogeneity and addressing price endogeneity. The framework has been used extensively to estimate demand in various markets/industries, which in turn provides bases for analyzing market outcomes and policy issues.

Recent breakthroughs in digital technology make a vast amount of data available for raw characteristics of the products, which makes, in practice, estimation of these models face several challenges. First, the standard estimation procedure, nested fixed point GMM, is computationally intensive and can be numerically unstable (see discussions in Knittel and Metaxoglou, 2012; Conlon and Gortmaker, 2020). Furthermore, largely because of this difficulty, researchers are restricted to imposed strong parametric assumptions on the distribution of random coefficients, e.g., normal distribution (almost exclusively used in practice), to reduce the number of parameters and thus to simplify the estimation problem. However, such restrictions are often not well motivated by economic theory and thus increase the risk of misspecification. Finally, given the inherent non-linearity of the model, it is difficult to pinpoint the fundamental variation in the data that drives estimates of substitution patterns in applications, which gives rise to the weak instruments problem (see e.g., Reynaert and Verboven, 2014; Gandhi and Houde, 2019).

Alternatively, flexible demand estimation models based on the identification argument of Berry and Haile (2014) have been recently proposed. Compiani (2018) follows the framework of Berry and Haile (2014) and demonstrates the performance of the NPIV estimator in a very simple case of two products with two characteristics. He uses Bernstein polynomials along with shape restrictions to alleviate the curse of dimensionality and nonparametrically estimate the inverse demand function. Methodology developed by Gandhi et al. (2020), hereafter GNT, is complementary, and allows the practitioner to apply it to more realistic settings. They resort to the dimensionality reduction idea of Gandhi and Houde (2019) which mitigates the curse of dimensionality and allows them to stay within the standard NPIV framework. Lu et al. (2019) consider a similar framework to GNT, but they focus on applications with large amounts of products instead of large amount of markets. How-

ever, both approaches still break down when the characteristics space becomes moderate- and/or high-dimensional. In attempt to address high-dimensionality in the nonparametric environment, Bakhitov et al. (2020) assume that consumer choices depend on a small set of product "features", which can be represented by some possibly nonlinear transformations of product characteristics, implying a sparsity condition on the true data generating process. Fosgerau et al. (2020) and Monardo (2021) consider a different class of inverse product differentiation models which generalize the inverse demand function of the nested logit model.

The remainder of the Chapter is organized as follows. In Section 3.2 we introduce the nonparametric demand estimation framework and discuss identification. In Section 3.3 we discuss the conditional demand derivative functional and how it can be used to estimated substitution patterns. Section 3.4 demonstrates the performance of the debiasing procedure from Chapter 2 applied to the conditional demand derivative. Section 3.5 estimates substitution patterns in the market for sodas using scanner data. Section 3.6 concludes.

3.2. Model and estimation framework

In this Section, we introduce a new framework for demand estimation that follows Gandhi et al. (2020) (hereafter, GNT). GNT is a flexible framework that combines the nonparametric identification arguments of Berry and Haile (2014) with the dimensionality reduction techniques of Gandhi and Houde (2019), which makes it applicable to real data sets with more than two products unlike Compiani (2018) whose approach fails due to the curse of dimensionality.

We follow Berry and Haile (2014) and present a general model of demand first, later on we will impose additional restrictions on the form of the indirect utility function as in GNT. In market t, $t = 1, \ldots, T$, there is a continuum of consumers choosing from a set of products $\mathcal{J} = \{0, 1, \ldots, J\}$ which includes the outside option. The choice set in market t is characterized by a set of product characteristics χ_t partitioned as follows:

$$\chi_t \equiv (x_t, p_t, \xi_t),$$

where $x_t \equiv (x_{1t}, \ldots, x_{Jt})$ is a vector of exogenous observable characteristics (e.g. exogenous product characteristics or market-level income), $p_t \equiv (p_{1t}, \ldots, p_{Jt})$ are observable endogenous characteristics (typically, market prices) and $\xi_t \equiv (\xi_{1t}, \ldots, \xi_{Jt})$ represent unobservables potentially correlated with p_t (e.g. unobserved product quality). Let $\mathcal{X}$ denote the support of χ_t. Then the structural demand system is given by

$$\sigma : \mathcal{X} \mapsto \Delta^J,$$

where Δ^J is a unit J-simplex. The function σ gives, for every market t, the vector s_t of shares for the J goods.

Following Berry and Haile (2014), we partition the vector of exogenous characteristics as $x_t = \left(x_t^{(1)}, x_t^{(2)}\right)$, where $x_t^{(1)} \equiv \left(x_{1t}^{(1)}, \ldots, x_{Jt}^{(1)}\right)$, $x_{jt} \in \mathbb{R}$ for $j \in \mathcal{J}\backslash\{0\}$, and define the linear indices

$$\delta_{jt} = x_{jt}^{(1)}\beta_j + \xi_{jt}, \quad j \in \mathcal{J}\backslash\{0\},$$

and let $\delta_t \equiv (\delta_{1t}, \ldots, \delta_{Jt})$. Without loss of generality, we can normalize $\beta_j = 1$ for all j (see Berry and Haile (2014) for more details). Given the definition of the demand system, for every market t,

$$\sigma(\chi_t) = \sigma\left(\delta_t, p_t, x_t^{(2)}\right).$$

Following Berry et al. (2013) and Berry and Haile (2014), we can show that there exists at most one vector δ_t such that $s_t = \sigma\left(\delta_t, p_t, x_t^{(2)}\right)$, meaning that we can write

$$\delta_{jt} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right), \quad j \in \mathcal{J}\backslash\{0\}. \tag{3.1}$$

We can rewrite (3.1) in a more convenient form to get the following estimation equation

$$x_{jt}^{(1)} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) - \xi_{jt}. \tag{3.2}$$

Note that in (3.2) the inverse demand is indexed by j, meaning that we have to estimate J inverse demand functions, that is exactly why the approach of Compiani (2018) gets severely hit by the curse of dimensionality. To circumvent this problem, Gandhi and Houde (2019) suggest transforming the input vector space under the linear utility specification to get rid of the j subscript. GNT follow this idea and show that Equation (3.2) can be rewritten as

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = x_{jt}^{(1)} + \gamma(\omega_{jt}) + \xi_{jt}, \tag{3.3}$$

where γ is such that

$$\sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) = \log\left(\frac{s_{jt}}{s_{0t}}\right) - \gamma(\omega_{jt}),$$

and $\omega_{jt} \equiv (\{s_{kt}, \Delta_{jkt}\}_{j\neq k})$, where $\Delta_{jkt} = \tilde{x}_{jt} - \tilde{x}_{kt}$ and $\tilde{x}_t \equiv \left(p_t, x_t^{(2)}\right)$.

Let $y_{jt} \equiv \log(s_{jt}/s_{0t}) - x_{jt}^{(1)}$, then we can rewrite equation (3.3) in a more convenient form

$$y_{jt} = \gamma(\omega_{jt}) + \xi_{jt}. \tag{3.4}$$

Equation (3.4) is the main structural equation where γ is a complex non-parametric function characterizing the relationship between the inverse demand and product attributes and shares. Dimensionality of the input vector ω_{jt} depends on both the dimensionality of the characteristics space and the number of products in the market, thus, ω_{jt} is potentially high-dimensional. This will always be the case if we want to augment standard datasets with unstructured data such as product reviews, package images, etc. Since both the market shares s_t and prices p_t depend on the unobservable characteristics ξ_t, $\mathbb{E}[\xi_{jt}|\omega_{jt}] \neq 0$, and hence, ω_{jt} is endogenous.

In order to estimate γ, we need to construct a vector of instruments z_{jt}. Berry et al. (1995) argue that the vector of product characteristics x_{jt} is exogenous with respect to the structural error term ξ_{jt}, i.e. $\mathbb{E}[\xi_{jt}|x_{jt}] = 0$. This exogeneity condition can be used to construct demand side instruments z_{jt}. Instrument construction is a well-known problem in demand estimation, since it can lead to weak identification and distorted inference. We refer the reader to Reynaert and Verboven (2014) and Gandhi and Houde (2019) for a more detailed discussion.

To construct demand side instruments, we follow Gandhi and Houde (2019) and use the transformed characteristics space $z_{jt} = (\{\Delta^x_{jkt}\}_{j\neq k})$, where $\Delta^x_{jkt} = x_{jt} - x_{kt}$, such that $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$. Note that since ω_{jt} includes z_{jt}, it enforces strong correlation between endogenous inputs and instruments. If data permit, one can augment the instrument space with supply side instruments, such as cost shifters. Let c_{jt} be a cost shifter for product j in market t, then the instrument space becomes $z_{jt} = \left(\left\{\Delta^x_{jkt}, \Delta^c_{jkt}\right\}_{j\neq k}\right)$, where $\Delta^c_{jkt} = c_{jt} - c_{kt}$.

3.3. Conditional demand function

One of the main primitives in demand estimation is substitution patterns which allow the researcher to investigate the responsiveness of consumer choices to changes in the market structure and, thus, understand the nature of competition between firms. Traditional metrics used to evaluate substitution patterns are price elasticities and diversion ratios. The price elasticity of product j to a price change in product k measures how demand for product j changes with the corresponding change in the price of product k. The diversion ratio for products j and k is the fraction of consumers who leave product j after a price increase and switch to product k. Both of those measures are widely used in industrial organization and anti-trust literature.

However, the nonparametric demand estimation framework, and especially the GNT framework, provide us with a novel object that can be used to measure substitution patterns.

Recall equation (3.3),

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = \underbrace{x_{jt}^{(1)} + \gamma(\omega_{jt}) + \xi_{jt}}_{\text{conditional demand}},$$

where the right-hand side of the expression above can be seen as a conditional demand function for product j in market t. The conditional demand function characterizes the relationship between the demand for product j (or the logarithm of the ratio of the share of product j to the share of the outside good) and product characteristics given shares of other products in the market.

In the GNT framework, the conditional demand function is the main building block for measuring substitution patterns. Let us rewrite equation (3.3) as

$$\Upsilon_{jt} \equiv \log\left(\frac{s_{jt}}{s_{0t}}\right) - x_{jt}^{(1)} - \gamma(\omega_{jt}) - \xi_{jt} = 0.$$

Let $\Upsilon_t \equiv (\Upsilon_{1t}, \ldots, \Upsilon_{Jt})$, then by the implicit function theorem, the gradient of the share vector in market t with respect to the vector of prices is given by

$$\nabla_{p_t} s_t = -[\nabla_{s_t} \Upsilon_t]^{-1} \nabla_{p_t} \Upsilon_t.$$

Note that $\nabla_{p_t} s_t$ depends on the gradients of the conditional demand function with respect to shares and prices.

For the rest of the paper we will focus on the conditional demand derivative with respect to own price. Note that this derivative is simply equal to $\partial\gamma(\omega_{jt})/\partial p_{jt}$. This object has a nice economic interpretation and connections to traditional parametric demand estimation models such as logit and nested logit, which we explore in greater detail in the following subsection.

Let $W_{jt} \equiv (y_{jt}, \omega_{jt}, z_{jt})$ be a data tuple. We use θ_{jk} to denote the conditional demand derivative functional such that

$$\theta_{jk} = \mathbb{E}[m(W_{jt}, \gamma)] = \mathbb{E}\left[\frac{\partial}{\partial p_{kt}} \gamma(\omega_{jt})\right] = \mathbb{E}[\alpha_{jk}(z_{jt})\gamma(\omega_{jt})],$$

where α_{jk} is the Riesz representer labeled by jk, meaning that for each product pair we have to estimate its corresponding Riesz representer. We can construct the debiased estimator

for θ_{jk} and its associated asymptotic variance using the formulae in (2.10),

$$\hat{\theta}_{jk} = \frac{1}{T}\sum_{\ell}^{L}\sum_{t\in T_\ell}\{m(W_{jt},\hat{\gamma}_\ell) + \hat{\alpha}_{jk,\ell}[y_{jt} - \hat{\gamma}_\ell(\omega_{jt})]\}$$

$$\hat{V}_{jk} = \frac{1}{T}\sum_{\ell}^{L}\sum_{t\in T_\ell}\hat{\psi}^2_{jk,\ell}(W_{jt}), \quad \hat{\psi}_{jk,\ell}(W_{jt}) = m(W_{jt},\hat{\gamma}_\ell) - \hat{\theta}_{jk} + \hat{\alpha}_{jk,\ell}[y_{jt} - \hat{\gamma}_\ell(\omega_{jt})].$$

Note that in the expressions above we treat one market as one observation, hence, the cross-fitting is applied across markets.

3.4. Simulated data experiments

3.4.1. Logit model

We focus on the derivative of the conditional demand function for good j with respect to its own price, $\theta_{jj} = \mathbb{E}[\partial\gamma(\omega_{jt})/\partial p_{jt}]$. This derivative measures sensitivity of the logarithm of the shares ratio to changes in price of product j conditional on the shares of competing products. When we fix the shares of other products in the market, the only two quantities that can change on the left-hand side of (3.3) in response to a price change are s_{jt} and s_{0t}. Thus, changes in s_{jt} can only occur at the expense of the corresponding change is s_{0t}. For example, if θ_{jj} is negative, it means that an increase in p_{jt} leads to a decrease in s_{jt} and a corresponding increase in s_{0t}, implying a positive substitution effect toward the outside good.

To get a better understanding of the interpretation, let us consider a simple logit model. The logit estimation equation takes the form

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = \beta_p p_{jt} + x'_{jt}\beta_x + \xi_{jt},$$

where $x_{jt} = \left(x^{(1)}_{jt}, x^{(2)}_{jt}\right)$. Recall, the GNT estimation equation is

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = x^{(1)}_{jt} + \gamma(\omega_{jt}) + \xi_{jt}.$$

Thus, if we take the mean derivative of γ with respect to own price, it will correspond to the price coefficient in the logit model, $\theta_{jj} = \beta_p$, given the data are coming from the logit model. In the logit case, the conditional and unconditional demand functions coincide, hence, the price derivative does not depend on the shares of competing products. We use this observation for our next simulation exercise.

We simulate data from a simple logit model. The mean valuation in the logit model is given by $\delta_{jt} = \beta_p p_{jt} + x'_{jt}\beta_x + \xi_{jt}$. Product shares can be calculated using the following formulae, for $j = 1, \ldots, J$ and $t = 1, \ldots, T$,

$$s_{jt} = \frac{\exp(\delta_{jt})}{1 + \sum_{j=1}^{J} \exp(\delta_{jt})} \quad \text{and} \quad s_{0t} = \frac{1}{1 + \sum_{j=1}^{J} \exp(\delta_{jt})}.$$

We set the total number of product characteristics besides the price to be equal to 4, i.e. $x_{jt}^{(1)}$ is a scalar and $x_{jt}^{(2)}$ is a three-dimensional vector. We draw the observed product characteristics, x_{jt}, from the standard normal distribution, while the unobserved characteristics, ξ_{jt}, are distributed as $\mathcal{N}(0, 0.25)$ for all j and t. The price is

$$p_{jt} = 2\left| x_{jt}^{(1)} + \sum_{k=1}^{3} x_{k,jt}^{(2)} + c_{jt} + \xi_{jt} + e_{jt} \right|,$$

where $c_{jt} \sim \mathcal{N}(0, 1)$ is a cost-shifter and $e_{jt} \sim \mathcal{N}(0, 0.01)$ is some additional noise. The price coefficient is $\beta_p = -2$ and the coefficients on product characteristics are $\beta_x = (1, -0.5, 0.5, 1)'$.

We use KIV[14] with the Gaussian RBF kernel to estimate γ. We construct dictionaries $b(z_{jt})$ and $d(\omega_{jt})$ using quadratic polynomials with interactions. Under the specified DGP, $\omega_{jt} = (\{s_{kt}, \Delta_{jkt}\}_{j \neq k})$ and $z_{jt} = \left(\left\{\Delta_{jkt}^{x}, \Delta_{jkt}^{c}\right\}_{j \neq k}\right)$, and hence, $dim(\omega_{jt}) = dim(z_{jt})$ and $p = q$. We run 200 replications for $J = 4, 6, 8$ and $T = 100, 200, 400$. We use five-fold cross-fitting, $L = 5$.

Table 5: MC results: logit price coefficient.

		PI Bias	DB Bias	PI SD	DB SD	PI RMSE	DB RMSE	PI Cvg	DB Cvg
$J = 4$	$T = 100$	0.691	0.133	0.225	0.485	0.727	0.503	0.00	0.95
	$T = 200$	0.500	0.199	0.091	0.243	0.508	0.314	0.00	0.76
	$T = 400$	0.466	0.239	0.079	0.159	0.473	0.287	0.00	0.56
$J = 6$	$T = 100$	0.376	0.088	0.058	0.341	0.380	0.352	0.00	0.96
	$T = 200$	0.311	0.048	0.040	0.316	0.313	0.320	0.00	0.93
	$T = 400$	0.293	0.079	0.032	0.149	0.295	0.169	0.00	0.92
$J = 8$	$T = 100$	0.262	0.045	0.042	0.092	0.265	0.102	0.00	0.97
	$T = 200$	0.212	0.008	0.029	0.061	0.214	0.062	0.00	0.93
	$T = 400$	0.181	0.002	0.024	0.041	0.183	0.041	0.00	0.92

Without loss of generality, we focus on the derivative of the conditional demand function for product 1. Table 5 presents the results. We can clearly see the bias-variance trade-

[14]Code: https://github.com/r4hu1-5in9h/KIV

off: the plug-in estimator has a much higher bias and smaller variance than the debiased estimator. This results in an extremely poor coverage of the plug-in estimator, which is essentially zero in all cases. Debiasing not only helps to diminish the bias, but also corrects the variance by adding the variation in the influence function to ensure proper coverage. Despite higher variance, the debiased estimator still has a lower RMSE. The coverage of the debiased estimator is close to the nominal 95% level across almost all specifications. For $J = 4$ and $T = 200, 400$ the debiasing is not that prominent which results into worse coverage compared to other specifications.

3.4.2. Nested logit model

Another model that has a closed form conditional demand function is nested logit. The estimation equation is given by

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = \beta_p p_{jt} + x'_{jt}\beta_x + \pi \log(s_{j|gt}) + \xi_{jt},$$

where $s_{j|gt}$ is the within nest share of product j in group g and $\pi \in [0, 1]$ characterizes the correlation of utilities that a consumer experiences among the products in the same nest. For simplicity, we assume that there are two mutually exclusive nests, $g = 1, 2$, and the outside good belongs to neither of the nests. Unlike logit, the conditional demand function under the nested logit model is different from the unconditional demand function. It implicitly depends on shares of within group products as $s_{j|gt} = 1 - \sum_{k \neq j,\, k \in \mathcal{J}_g} s_{k|gt}$, where $\mathcal{J}_g$ denotes products that belong to group g.

Under the nested logit specification, the derivative of the conditional demand function with respect to price takes the following form

$$\theta_{jj} = \mathbb{E}\left[\beta_p + \frac{\pi}{s_{j|gt}} \frac{\partial s_{j|gt}}{\partial p_{jt}}\right]. \tag{3.5}$$

To proceed, let us first focus on the derivative of the within group share with respect to the mean valuation, i.e. $\partial s_{j|gt}/\partial \delta_{jt}$, which is given by

$$\frac{\partial s_{j|gt}}{\partial \delta_{jt}} = \frac{1}{1-\pi} s_{j|gt}(1 - s_{j|gt}).$$

Applying the chain rule,

$$\frac{\partial s_{j|gt}}{\partial p_{jt}} = \frac{\partial s_{j|gt}}{\partial \delta_{jt}} \frac{\partial \delta_{jt}}{\partial p_{jt}} = \frac{\beta_p}{1-\pi} s_{j|gt}(1 - s_{j|gt}). \tag{3.6}$$

Thus, combining (3.5) and (3.6) gives

$$\theta_{jj} = \mathbb{E}\left[\beta_p\left(1 + \frac{\pi}{1-\pi}(1 - s_{j|gt})\right)\right] = \underbrace{\beta_p}_{\text{direct effect}} + \underbrace{\beta_p \frac{\pi}{1-\pi}\mathbb{E}[1 - s_{j|gt}]}_{\text{indirect effect}}. \tag{3.7}$$

Note that in equation (3.7) the effect of a price change comes from two components. The direct effect measures how changes in p_{jt} affect the log ratio of the shares, which is similar to the logit price coefficient. The indirect effect measures the effect price changes have through the nesting structure. Note that the closer the nesting parameter is to 1, the larger the derivative becomes. When π is close 1, consumers tend to substitute more towards products within the nest. As a result, conditional on the shares of other within group products, consumers prefer to substitute more towards the outside good rather than towards the off-nest products, hence, the derivative is larger in absolute value. On the other hand, the larger the within group share of product j is, the smaller is the derivative. If product j has a large share within group g, it means that consumers strongly prefer product j to other products in the nest, and hence, less sensitive they are to its price changes.

Since nested logit enforces stronger substitution effects between products from the same nest, the shares formulae are more complicated than in the logit case. If product j belongs to group g, then the choice probability of product j in market t conditional on group g being chosen equals

$$s_{j|gt} = \frac{\exp\left(\frac{\delta_{jt}}{1-\pi}\right)}{D_{gt}}, \quad \text{where} \quad D_{gt} = \sum_{j \in \mathcal{J}_g} \exp\left(\frac{\delta_{jt}}{1-\pi}\right).$$

The probability that group g is chosen equals

$$s_{gt} = \frac{D_{gt}^{1-\pi}}{\sum_g D_{gt}^{1-\pi}}.$$

Thus, the unconditional probability of product j, from nest g, in market t being chosen is given by

$$s_{jt} = s_{j_g t} s_{gt} = \frac{\exp\left(\frac{\delta_{jt}}{1-\pi}\right)}{D_{gt}^{\pi}\left[\sum_g D_{gt}^{1-\pi}\right]}.$$

To see whether ML is capable of capturing this departure from logit, we run another Monte Carlo exercise. Similarly to the logit design, we set the number of observed product characteristics to 4. However, now there is one categorical characteristic which defines nests and

does not change across markets, $x^{(2)}_{3,jt}$. It assigns the first half of the products in the market to the first group and the second half to the second group. For example, if $J = 4$, then products 1 and 2 belong to the first nest, while products 3 and 4 belong to the second nest. The remaining product characteristics are drawn from the standard normal distribution. All the remaining quantities are constructed in the same fashion as in the logit design. We set $\pi = 0.5$. The estimation procedure is unchanged.

Table 6: MC results: nested logit own price derivative.

		PI Bias	DB Bias	PI SD	DB SD	PI RMSE	DB RMSE	PI Cvg	DB Cvg
$J = 4$	$T = 100$	0.837	0.097	0.458	0.491	0.954	0.501	0.01	0.90
	$T = 200$	0.609	0.001	0.240	0.310	0.655	0.310	0.00	0.88
	$T = 400$	0.496	0.018	0.168	0.220	0.524	0.221	0.01	0.90
$J = 6$	$T = 100$	0.455	0.033	0.201	0.245	0.498	0.247	0.01	0.82
	$T = 200$	0.297	0.058	0.124	0.162	0.321	0.172	0.02	0.75
	$T = 400$	0.231	0.032	0.112	0.132	0.257	0.136	0.06	0.71
$J = 8$	$T = 100$	0.309	0.040	0.134	0.183	0.337	0.188	0.04	0.81
	$T = 200$	0.200	0.007	0.084	0.118	0.217	0.118	0.06	0.81
	$T = 400$	0.123	0.034	0.086	0.092	0.150	0.091	0.13	0.78

The results are presented in Table 6. The overall pattern is similar to the logit case. However, the bias-variance trade-off is not that prominent in the nested logit case. The variance of the debiased estimator is still higher than of the plug-in estimator, however, the gap becomes much smaller. Moreover, unlike the logit case, for $J = 4$ debiasing works equally well across all sample sizes. Finally, despite the debiased estimator achieves much better coverage than the plug-in estimator, it still undercovers in all specifications.

3.5. Estimation of substitution patters in the market for sodas

We use retail scanner data from the IRI Academic Database (Bronnenberg et al., 2008). This dataset includes unit sales by UPC code, store and week for a sample of supermarkets over 2001-2012 as well as information on product characteristics. We focus on one year span of 2003 and top ten most sold products. Among others, the list of products includes Coke, Pepsi, Sprite, Dr. Pepper, etc. Since we want to exploit the variation in product attributes, we do not aggregate the data to the brand level. In other words, products are defined by a combination between a brand and a set of product characteristics.

Carbonated beverages are sold in different packages and package sizes. We restrict our attention to cans and define a product unit as a 12 oz can. Hence, we construct market shares based on the total amount of cans sold. Prices are defined as the ratio of total revenue to total number of units sold. We aggregate the data to geographic region-month level resulting in 5,640 observations at the product-region-month level.

Data on product characteristics include beverage flavor, sugar, caffeine and calorie levels. All characteristics are represented by categorical variables, thus, for computational reasons we aggregate product attributes in larger groups (see Appendix C.1 for more details). After aggregation and dropping collinear characteristics, we are left with six product attributes we use for estimation. We have CAFFEINE and SUGAR dummy variables indicating whether a product contains caffeine and sugar, respectively. The remaining four variables represent different flavor categories: COLA, LEMONADE, PEPPER, and OTHERS.

We start our analysis with parametric specifications. To be precise, we estimate logit, nested logit, and BLP models. Since beverage flavors are represented by four dummy variables, we drop the intercept to avoid collinearity issues. For the nested logit specification, we split the products into two categories based on the amount of sugar. When estimating the BLP model, we put random coefficients on price, CAFFEINE, and SUGAR, while keeping the flavor variables only in the linear part. We also consider two sets of instruments: (i) standard BLP instruments and (ii) local differentiation instruments of Gandhi and Houde (2019). All models are estimated with the PyBLP package (Conlon and Gortmaker, 2020) available in `Python`.

Table 7 displays the results. We can observe that the linear coefficients estimates are pretty close across the estimators. Positive coefficients on CAFFEINE and SUGAR imply that consumers tend to prefer beverages containing caffeine and sugar over decaffeinated and diet alternatives. As we have dropped the intercept term, we can only interpret differences in flavor dummies. As COLA has the largest coefficient among all other flavors, we conclude that consumers prefer cola-flavored drinks over other alternatives.

The price coefficient in the nested logit is slightly smaller compared to the logit estimate since a part of the price effect comes through the within group share, which is captured in (3.7). The nesting coefficient estimate equals to 0.195 indicating a relatively weak nesting structure. BLP specifications mostly differ in the estimates of nonlinear parameters. Using the vanilla BLP instruments uncovers more heterogeneity in consumer preferences across sugar levels, while using the differentiation IVs picks up more heterogeneity across caffeine levels.

Overall, the price coefficient estimates give us a sense of the order of magnitude of the conditional demand function derivative. As in the simulated data experiments, we use KIV to estimate the conditional demand function γ. Besides prices and shares, there are 5 product characteristics in $x_{jt}^{(2)}$ leading to $dim(\omega_{jt}) = 70$, which makes the problem moderately high-dimensional. To construct $b(z_{jt})$ and $b(\omega_{jt})$ dictionaries, we use empirical moment based

Table 7: Parametric demand estimates

	Logit	Nested Logit	BLP	BLP DIV*
		Linear parameters		
price	-5.353	-4.564	-5.369	-5.364
	(0.060)	(0.004)	(0.071)	(0.064)
CAFFEINE	0.522	0.409	0.522	0.414
	(0.051)	(0.046)	(0.056)	(0.052)
SUGAR	0.735	0.710	0.095	0.710
	(0.066)	(0.061)	(0.065)	(0.073)
COLA	-1.491	-1.363	-1.496	-1.495
	(0.032)	(0.031)	(0.035)	(0.033)
LEMONADE	-2.441	-2.133	-2.447	-2.450
	(0.059)	(0.045)	(0.073)	(0.064)
PEPPER	-2.630	-2.258	-2.643	-2.638
	(0.078)	(0.057)	(0.092)	(0.082)
OTHERS	-3.840	-3.182	-3.856	-3.849
	(0.052)	(0.036)	(0.060)	(0.054)
		Nonlinear parameters		
$\hat{\pi}$	–	0.195	–	–
		(0.013)		
price	–	–	0.461	0.332
			(0.012)	(0.008)
CAFFEINE	–	–	0.145	1.168
			(0.005)	(0.010)
SUGAR	–	–	2.191	0.413
			(0.030)	(0.006)

* Estimated using local differentiation instruments.

basis functions as in GNT (see Appendix C.2 for more details) with $p = 405$ and $q = 594$. The debiased estimator is constructed using five-fold cross-fitting, $L = 5$. We compare the performance of the debiased estimator to the plug-in KIV and nested logit estimators.

Table 8 presents conditional demand function derivative estimates for each product. The first two columns contain nested logit estimates and their corresponding standard errors. Nested logit estimates are constructed by simply plugging-in the estimated parameters into (3.7) and replacing the expectation with the sample average. There is no much variation in the estimates across products with estimated values being close the logit price coefficient estimate. Unlike the nested logit estimates, the KIV plug-in estimates do exhibit substantial variation across products. Moreover, we can observe that products with similar characteristics exhibit similar responsiveness to price changes. For example, diet Coke and diet Pepsi have similar derivative estimates, while regular Coke and regular Pepsi are more sensitive. Given the categorical nature of the characteristics space, we can interpret this findings as

uncovering the underlying nesting structure in the data.

Table 8: Conditional demand derivative estimates.

Product	NL	NL SE	PI	PI SE	DB	DB SE
Caffeine-free Diet Coke	-5.502	0.085	-3.453	0.061	-4.375	0.314
Caffeine-free Diet Pepsi	-5.575	0.081	-3.364	0.063	-4.420	0.332
Coke	-5.262	0.188	-5.094	0.032	-6.049	0.249
Pepsi	-5.340	0.193	-5.046	0.045	-6.172	0.302
Diet Coke	-5.169	0.218	-4.034	0.021	-4.667	0.171
Diet Pepsi	-5.327	0.210	-4.041	0.025	-4.800	0.204
Dr. Pepper	-5.572	0.122	-3.106	0.082	-5.635	0.693
Mountain Dew Classic	-5.561	0.086	-4.585	0.066	-6.863	0.617
Mountain Dew Other	-5.637	0.058	-2.954	0.103	-5.784	0.781
Sprite	-5.539	0.060	-4.015	0.020	-6.025	0.552

We can see a clear debiasing effect in the last two columns of Table 8. First, the plug-in KIV estimates are biased upwards. It is important to note that despite being numerically different the debiased estimates are qualitatively close to the plug-in estimates and preserve the data patterns uncovered by the KIV estimator. This indicates that debiasing indeed corrects for the regularization bias without distorting the estimates. Second, the standard errors after debiasing are larger than those of the plug-in estimator. These findings are coherent with the Monte Carlo evidence from Chapter 2.

3.6. Conclusion

In this Chapter, we showed how to apply the debiasing procedure from Chapter 2 to estimate the conditional demand derivative in the nonparametric demand for differentiated products framework. We have obtained evidence from both simulated and real scanner data that plug-in estimates are biased upwards and have smaller variance compared to the debiased estimates, which reflects the bias-variance trade-off occurring due to regularization. Looking at the conditional demand derivative is a first step towards understanding the benefits of using machine learning to estimate substitution patterns over the standard parametric methods. Thus, taking one step further to estimation of classical measures of substitution like elasticities and diversion ratios and to counterfactual analysis seems like a natural addition to the future research agenda.

APPENDIX A: Additional Details and Proofs for Chapter 1

A.1. Auxiliary lemmas

Lemma A.1.1. Assume the assumptions of Lemma 1 are satisfied. Consider h_m that satisfies Assumption 3. Let $\bar{f}$ be an arbitrary reference function in $\mathcal{S}$. Also, define $s_m = ||f_0||_1 + \sum_{i=0}^{m-1} h_i$ and

$$\Delta\hat{Q}(f) = \max\left(0, \hat{Q}(f) - \hat{Q}(\bar{f})\right), \tag{A.1.1}$$

$$\bar{\epsilon}_m = \frac{h_m^2}{2}M + \epsilon_m. \tag{A.1.2}$$

Then after m steps the following bound holds for f_{m+1},

$$\Delta\hat{Q}(f_{k+1}) \leq \left(1 - \frac{h_m}{s_m + ||\bar{f}||_1}\right)\Delta\hat{Q}(f_m) + \bar{\epsilon}_m. \tag{A.1.3}$$

Proof. The result follows directly from Lemma 1 and Lemma 4.1 in Zhang and Yu (2005). ■

Lemma A.1.2. Under the assumptions of Lemma A.1.1, we have

$$\Delta\hat{Q}(f_m) \leq \frac{||f_0||_1 + ||\bar{f}||_1}{s_m + ||\bar{f}||_1}\Delta\hat{Q}(f_0) + \sum_{j=1}^{m}\frac{s_j + ||\bar{f}||_1}{s_m + ||\bar{f}||_1}\bar{\epsilon}_{j-1}. \tag{A.1.4}$$

Proof. The above lemma directly follows from the repetitive application of Lemma A.1.1. For detailed proof see Zhang and Yu (2005). ■

Lemmas A.1.1 and A.1.2 are direct counterparts of Lemmas 4.1 and 4.2 in Zhang and Yu (2005) with $M(s_{m+1})$ replaced by M. Therefore, the main numerical convergence result below follows as well (see Corollary 4.1).

A.2. Proofs of results

A.2.1. Proof of Lemma 1

First, $Q(\cdot)$ is convex in f, hence, it is convex differentiable. Now we have to bound the second derivative with respect to h. Note that the second derivative of $Q_{f,\varphi}(h)$ does not

even depend on h,

$$\begin{aligned} Q''_{f,\varphi}(h) &= \mathbb{E}[\varphi(x_i)z_i]'\Omega\mathbb{E}[\varphi(x_i)z_i] \\ &\leq \lambda_{max}(\Omega)||\mathbb{E}[\varphi(x_i)z_i]||^2 \\ &\leq \lambda_{max}(\Omega)\mathbb{E}[|\varphi(x_i)|^2]\mathbb{E}[|z_i'z_i|] \\ &\leq \lambda_{max}(\Omega)CB \equiv M < \infty, \end{aligned}$$

where the second inequality is a by the Cauchy-Schwarz inequality, and the last inequality comes from the assumptions of the lemma. Thus, the second derivative has a fixed bound $M < \infty$. ■

A.2.2. Proof of Theorem 1

The result follows directly from Lemmas A.1.1 and A.1.2. For detailed proof see Zhang and Yu (2005). ■

A.2.3. Proof of Lemma 2

Follows directly from Lemma 4.3 in Zhang and Yu (2005). ■

A.2.4. Proof of Lemma 3

It follows from Lemma 2 and condition (1.17) that for all $j = 1, \ldots, k$,

$$\mathbb{E}_W \sup_{||f||_1 \leq \beta} |g_{0,j}(f) - \hat{g}_j(f)| \leq 2\gamma_j(\beta)\beta R(\mathcal{S}) \leq 2\gamma_j(\beta)\beta\frac{C_{\mathcal{S}}}{\sqrt{n}} = O(n^{-1/2}).$$

Thus, by Markov inequality,

$$\sup_{||f||_1 \leq \beta} |g_{0,j}(f) - \hat{g}_j(f)| \xrightarrow{p} 0, \quad j = 1, \ldots, k.$$

Since every coordinate of the sample moment function converges uniformly to its population analog, we can bound the norm as well

$$||g_0(f) - \hat{g}(f)|| = \left(\sum_{j=1}^{k} |g_{0,j}(f) - \hat{g}_j(f)|^2\right)^{1/2} \leq \sqrt{k}O_p(n^{-1/2}),$$

which combined with Markov inequality completes the proof. ■

A.2.5. Proof of Theorem 2

By the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned}\left|\hat{Q}(f) - Q(f)\right| &\leq \left|[\hat{g}(f) - g_0(f)]'\hat{\Omega}[\hat{g}(f) - g_0(f)]\right| + \left|g_0(f)'(\hat{\Omega} + \hat{\Omega}')[\hat{g}(f) - g_0(f)]\right| \\ &\quad + \left|g_0(f)'(\hat{\Omega} - \Omega)g_0(f)\right| \\ &\leq ||\hat{g}(f) - g_0(f)||^2||\hat{\Omega}|| + 2||g_0(f)||\,||\hat{g}(f) - g_0(f)||\,||\hat{\Omega}|| + ||g_0(f)||^2||\hat{\Omega} - \Omega||.\end{aligned}$$

Using Lemma 3, (ii), and (iv) and taking the supremum of both sides of the inequality completes the proof. ■

A.2.6. Proof of Theorem 3

By Theorem 2, the first term converges in probability to zero, and the second term converges to zero according to the arguments from the proof of Theorem 3.1 in Zhang and Yu (2005), which completes the proof. ■

A.3. Alternative bound on the Rademacher complexity

To derive an alternative bound on the Rademacher complexity, we introduce the following lemma (Massart's lemma).

Lemma A.3.1. For any $A \subseteq \mathbb{R}^n$, let $M = \sup_{a \in A} ||a||$. Then

$$\hat{R}(A) = \mathbb{E}_\sigma \left[\sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i\right] \leq \frac{M\sqrt{2\log|A|}}{n}.$$

This lemma can be applied to any finite class of functions.

Example 5. Consider a set of binary classifiers $\mathcal{H} \subseteq \{h : W \mapsto \{-1, 1\}\}$. Given a sample $W = (W_1, \ldots, W_n)$, we can take $A = \{h(W_1), \ldots, h(W_n) \,|\, h \in \mathcal{H}\}$. Then $|A| = |\mathcal{H}|$ and $M = \sqrt{n}$. Massart's lemma gives

$$\hat{R}(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{n}}.$$

In general, Massart's lemma can also be applied to infinite function classes with a finite shattering coefficient. Notice that Massart's finite lemma places a bound on the empirical Rademacher complexity that depends only on n data points. Therefore, all that matters

as far as empirical Rademacher complexity is concerned is the behavior of a function class on those data points. We can define the empirical Rademacher complexity in terms of the shattering coefficient.

Lemma A.3.2. Let $\mathcal{Y} \subset \mathbb{R}$ be a finite set of real numbers of modulus at most $C > 0$. Given a sample $W = (W_1, \ldots, W_n)$, the Rademacher complexity of any function class $\mathcal{H} \subseteq \{h : W \mapsto \mathcal{Y}\}$ can be bounded in terms of its shattering coefficient $s(\mathcal{H}, n)$ by

$$\hat{R}(\mathcal{H}) \leq C\sqrt{\frac{2\log s(\mathcal{H}, n)}{n}}.$$

Proof. Let $A = \{h(W_1), \ldots, h(W_n) \,|\, h \in \mathcal{H}\}$, then $M = \sup_{a \in A} ||a|| = C\sqrt{n}$ and $|A| = s(\mathcal{H}, n)$. Applying the Massart's lemma gives

$$\hat{R}(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(W_i)\right] \leq \frac{M\sqrt{2\log|\mathcal{H}|}}{n} = C\sqrt{\frac{2\log s(\mathcal{H}, n)}{n}}.$$

■

Note that we apply the Massart's lemma conditional on the sample, hence, we can use the same bound for $\hat{R}(\mathcal{H})$. We can loosen the bound by applying Sauer's lemma which says that $s(\mathcal{H}, n) \leq n^d$, where d is the VC dimension of $\mathcal{H}$. This simplifies the result of Theorem A.3.2 to

$$\hat{R}(\mathcal{H}) \leq C\sqrt{\frac{2d\log(n)}{n}} = O\left(\sqrt{\frac{\log(n)}{n}}\right). \tag{A.3.1}$$

The bound in (A.3.1) is valid for any class with finite VC dimension which is coherent with the results of Zhang and Yu (2005). However, the VC bound is slower that the bound in (1.17) by the factor or $\log(n)$ which appears in a lot of ML algorithms.

Note that the bound in (A.3.1) is still a valid bound for the main results to follow. It only affects the rate of convergence.

APPENDIX B: Additional Details and Proofs for Chapter 2

B.1. Performance of standard ML algorithms under endogeneity

In this example, we illustrate that standard ML algorithms such as Neural Networks fail to capture the structural function under endogeneity.

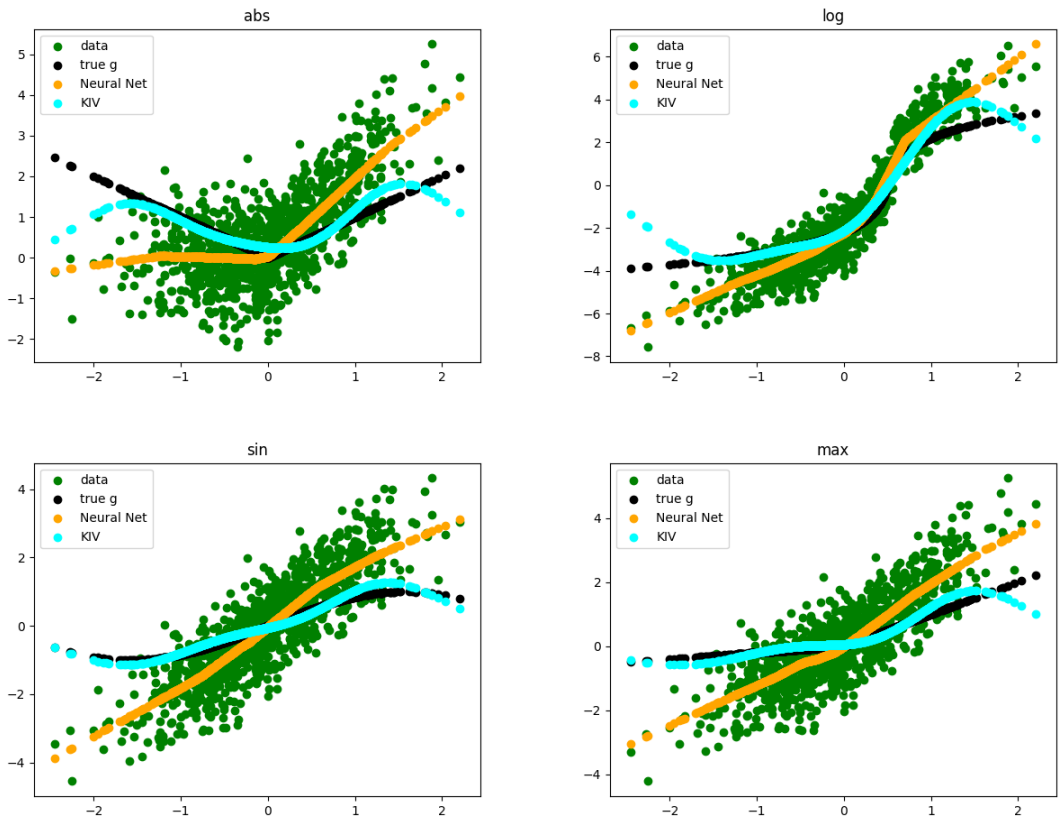

Figure 3: Standard ML vs MLIV estimators

Consider the following design similar to Lewis and Syrgkanis (2018), Bennett et al. (2019), and Bakhitov and Singh (2021). Let

$$Y_i = \gamma(X_i) + e_i + \delta_i, \quad X_i = 0.5Z_i + 0.5e_i,$$
$$Z_i \sim \mathcal{N}(0,1), \quad e_i \sim \mathcal{N}(0,1), \quad \delta_i \sim \mathcal{N}(0,0.1),$$

where e_i is a confounder. We consider four different choices of γ,

$$\textbf{abs: } \gamma(X) = |X|, \qquad \textbf{log: } \gamma(X) = \log(|16X - 8| + 1)\text{sign}(X - 0.5)$$
$$\textbf{sin: } \gamma(X) = \sin(X), \qquad \textbf{max: } \gamma(X) = \max(X, 0.2X).$$

We compare the performance of the standard 2-layer Neural Networks with $(16, 8)$ nodes to the performance of the Kernel IV regression of Singh et al. (2019). We use 2000 observations for training and 1000 observations for testing.

Figure 3 shows that the Neural Network fails to capture the structural function. We can see that in all cases it obviously is fitting the conditional expectation $\mathbb{E}[Y|X]$ instead of γ. In contrast, KIV is able to pick up the structural function in all cases, despite having problems at the boundaries which is a common problem of all kernel methods.

B.2. Analytical solution to the GMM problem

In this Section, we provide additional intuition behind the PGMM estimator of the RR. To do so, we focus on the standard GMM problem without adding the penalty term, i.e. $\hat{\rho}$ is a solution to

$$\min_{\rho \in \mathbb{R}^p} (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho). \tag{B.2.1}$$

Given the form of the debiased moment function (2.5) and the linear approximation for the RR, the orthogonal moment condition (2.7) will always be linear in ρ, meaning that the GMM criterion in (B.2.1) is globally concave and has a unique global minimizer.

For the ease of exposition, we drop the cross-fitting notation and assume that we are interested in a linear functional $\theta = \mathbb{E}[m(W, \gamma)]$. Then the moment condition takes the form

$$\hat{\psi}_\gamma(d_j, \rho) = \frac{1}{n}\sum_{i=1}^{n} \left\{ m(W_i, d_j) - d_j(X_i)b(Z_i)'\rho \right\}, \quad j = 1, \ldots, q.$$

Let $m(W, d) = (m(W, d_1), \ldots, m(W, d_q))$. Taking the first-order condition of the GMM criterion gives

$$\frac{\partial \hat{\psi}_\gamma(\hat{\rho})}{\partial \rho'}\hat{\Omega}_q \left\{ \frac{1}{n}\sum_{i=1}^{n} m(W_i, d) - \frac{1}{n}\sum_{i=1}^{n} d(X_i)b(Z_i)'\hat{\rho} \right\} = 0. \tag{B.2.2}$$

We can rewrite (B.2.2) in matrix form as

$$-\hat{G}'\hat{\Omega}_q\hat{M} + \hat{G}'\hat{\Omega}_q\hat{G}\hat{\rho} = 0,$$

which immediately gives a closed-form solution for $\hat{\rho}$,

$$\hat{\rho} = (\hat{G}'\hat{\Omega}_q\hat{G})^{-1}\hat{G}'\hat{\Omega}_q\hat{M}. \tag{B.2.3}$$

Note that the form of the GMM solution in (B.2.3) resembles the GMM solution to the classical linear IV problem, but with endogenous regressors and instruments being switched. Ichimura and Newey (2017) point out that $\alpha(Z)$ is the solution of a "reverse" structural equation involving an expectation conditional on the endogenous variables X rather than the instruments Z. If we set $\hat{\Omega} = \left(\frac{1}{n}\sum_{i=1}^{n} d(X_i)d(X_i)'\right)^{-1}$, we will get the exact solution to the "reverse" NPIV problem.

B.3. Computing Auto-DML using Penalized GMM

Recall, in matrix form the PGMM problem is given by

$$\min_{\rho\in\mathbb{R}^p} (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda_n|\rho|_1. \tag{B.3.1}$$

Note that the objective in (B.3.1) is a generalized version of the Lasso objective. Thus, we can generalize the coordinate decent approach for Lasso to the PGMM objective that we use in this paper. We follow CNS and use a coordinate-wise descent algorithm with the soft-thresholding update.

We denote the jth element of a generic vector v by v_j and let e_j be a $p \times 1$ unit vector with 1 in the jth coordinate and zeros elsewhere.

Algorithm 5 Coordinate-wise descent algorithm for PGMM

for $j = 1 : p$ **do**

Calculate loadings that do not depend on ρ_j:

$$B_j = e_j'\hat{G}'\hat{\Omega}_q\hat{G}e_j$$
$$A_j = e_j'\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho + \hat{G}e_j\rho_j)$$

Update coordinate ρ_j:

$$\rho_j = \begin{cases} \frac{A_j+\lambda_n}{B_j} & \text{if } A_j < -\lambda_n \\ 0 & \text{if } A_j \in [-\lambda_n, \lambda_n] \\ \frac{A_j-\lambda_n}{B_j} & \text{if } A_j > \lambda_n \end{cases}$$

end for

The justification for Algorithm 5 is similar to the one of CNS. It follows from the fact that

the GMM objective (B.3.1) is of the form of eq. 21 of Friedman et al. (2007), hence, the coordinate descent converges to the minimizer of the objective (Tseng, 2001).

One can boost the performance of the PGMM algorithm by incorporating adaptive penalty loadings in the spirit of Zou (2006). This will transform the optimization problem (B.3.1) into the adaptive PGMM (A-PGMM) problem

$$\min_{\rho} (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda_n \sum_{j=1}^{p} \hat{w}_j|\rho_j|, \tag{B.3.2}$$

where $\hat{w} = (\hat{w}_1, \dots, \hat{w}_p)$ is a vector of data-dependent weights with $\hat{w}_j = 1/|\tilde{\rho}_j|$, and $\tilde{\rho}$ is a preliminary consistent estimator. The only difference to Algorithm 5 is that in step 3 we replace λ_n with $\hat{w}_j\lambda_n$.

B.3.1. Numerical performance

We evaluate the numerical performance of the PGMM algorithm in two scenarios: (i) exogenous high-dimensional linear regression, (ii) high-dimensional linear IV regression.

HD linear regression

We borrow the set-up from CNS and compare the performance of PGMM and A-PGMM algorithms with the MD Lasso estimator of CNS as well as with the built-in `Python` implementations of the stochastic gradient descent (SGD) and least-angle regression (LARS) algorithms[15].

In this design, the data generating process is

$$Y = X'\beta_0 + \varepsilon,$$

where $X = (1, X_1, \dots, X_{100})'$, $X_j \sim \mathcal{N}(0,1)$ and i.i.d., and $\varepsilon \sim \mathcal{N}(0,1)$. The true value of the regression coefficient is $\beta_0 = (1,1,1,0,0,\dots)$ and $\dim(\beta_0) = 101$. The number of observations is $n = 100$. We can recover β_0 by using the functional $m(w,h) = yh(x)$ in the PGMM and MD Lasso formulations[16].

In Table 9, we report MSE defined as $|\hat{\beta} - \beta_0|_2^2$ of various implementations based on 200 simulations. We can see that PGMM performs on par with SGD, LARS, and MD Lasso, while

[15] We use `LassoCV` and `LassoLarsCV` commands to run SGD and LARS algorithms respectively.

[16] Alternatively, we could simply use the standard GMM moment $g(w,h) = (y - h(x))x$ for the linear regression to implement PGMM.

adding adaptive weights on the penalty term improves the performance twice rendering the lowest MSE across the algorithms, which validates the procedure.

Table 9: HD Linear regression results.

	MSE
SGD	0.1553
LARS	0.1786
MD Lasso	0.1474
PGMM	0.1791
A-PGMM	0.0868

HD linear IV regression

We follow the exponential design of Belloni et al. (2012). The DGP is

$$Y = X'\beta_0 + \varepsilon$$
$$X = \Pi Z + v,$$

where $\beta_0 = (1, 1, 1, 0, 0, \dots)$ and $\dim(\beta_0) = 101$, $X = (1, X_1, \dots, X_{100})'$, $Z = (Z_1, \dots, Z_{150}) \sim \mathcal{N}(0, \Sigma_Z)$ is a 150×1 vector with $\mathbb{E}[Z_j^2] = 1$ and $\text{Corr}(Z_h, Z_j) = 0.5^{|h-j|}$. We set the first stage coefficients $\Pi = (1, 0.7, 0.7^2, \dots, 0.7^{149})$. The structure of the error terms is the following: $\varepsilon \sim \mathcal{N}(0, 1)$ and $v|\varepsilon \sim \mathcal{N}(r\varepsilon, \mathcal{I} - r^2)$ so that the unconditional covariance matrix of the endogenous variables is the identity. We set $r = 0.5$ and the number of observations $n = 100$.

We compare the performance of PGMM and A-PGMM algorithms to the Double Lasso estimator of Gold et al. (2020). Table 10 demonstrates MSEs of the considered implementations based on 200 simulations.

Table 10: HD Linear IV regression results.

	MSE
Double Lasso	0.1864
PGMM	0.3020
A-PGMM	0.0726

B.4. Proofs of results

In this Section, we present the proofs of the theoretical results of the paper along with auxiliary lemmas and their corresponding proofs.

B.4.1. Properties of the PGMM estimator

Lemma B.4.1. If Assumption 6 is satisfied, then

$$||\hat{G} - G||_\infty = O_p(\varepsilon_n^G), \quad \varepsilon_n^G = \sqrt{\frac{\log(q)}{n}}.$$

Proof. The proof is similar to the proof of Lemma C1 of Chernozhukov, Newey and Singh (2018). Define

$$T_{ijk} = d_j(X_i)b_k(Z_i) - \mathbb{E}[d_j(X_i)b_k(Z_i)], \, U_{jk} = \frac{1}{n}\sum_{i=1}^{n} T_{ijk}.$$

For any constant C,

$$\begin{aligned}
\mathbb{P}(||\hat{G} - G||_\infty \geq C\varepsilon_n^G) &\leq \sum_{j=1}^{q}\sum_{k=1}^{p} \mathbb{P}(|U_{ijk}| \geq C\varepsilon_n^G) \\
&\leq pq \max_{j,k} \mathbb{P}(|U_{ijk}| \geq C\varepsilon_n^G) \\
&\leq q^2 \max_{j,k} \mathbb{P}(|U_{ijk}| \geq C\varepsilon_n^G),
\end{aligned}$$

where the last inequality follows from $q \geq p$. Note that $\mathbb{E}[T_{ijk}] = 0$ and by Assumption 6,

$$|T_{ijk}| \leq |d_j(X_i)| \cdot |b_k(Z_i)| + \mathbb{E}[|d_j(X_i)| \cdot |b_k(Z_i)|] \leq 2C_bC_d.$$

Since T_{ijk} is a bounded random variable, it is sub-Gaussian. Let $||T_{ijk}||_{\Psi_2}$ denote the sub-Gaussian norm. Define $K = 2C_bC_d/\log 2 \geq ||T_{ijk}||_{\Psi_2}$. By Hoeffding's inequality (see Theorem 2.6.2 in Vershynin, 2018), there is a constant c such that

$$\begin{aligned}
q^2 \max_{j,k} \mathbb{P}(|U_{ijk}| \geq C\varepsilon_n^G) &\leq 2q^2 \exp\left(-\frac{c(nC\varepsilon_n^G)^2}{nK^2}\right) \\
&= 2q^2 \exp\left(-\frac{cC^2\log(q)}{K^2}\right) \\
&\leq 2\exp\left(\log(q)\left[2 - \frac{cC^2}{K^2}\right]\right) \to 0
\end{aligned}$$

for any $C > K\sqrt{2/c}$. Thus, for large enough C, $\mathbb{P}(|\hat{G} - G|_\infty \geq C\varepsilon_n^G) \to 0$, which completes the proof. ■

Lemma B.4.2. For any $q \times 1$ vector $\hat{M}$, $q \times p$ matrix $\hat{G}$, $q \times q$ matrix $\hat{\Omega}$, and $\lambda > 0$, if

$$\rho^* = \underset{\rho \in \mathbb{R}^p}{\text{argmin}} \left\{ (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda|\rho|_1 \right\},$$

then

$$||\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho^*)||_\infty \leq \lambda.$$

Proof. The proof is similar to the proof of Lemma C0 of Chernozhukov, Newey and Singh (2018). Since the objective function is convex in ρ, a necessary condition for minimization is that zero belongs to the sub-differential of the objective function, i.e.

$$0 \in -\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho^*) + \lambda([-1,1], \ldots, [-1,1])'.$$

Thus, for $j = 1, \ldots, p$ we have

$$-e_j'\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho^*) + \lambda \geq 0, \quad -e_j'\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho^*) - \lambda \leq 0,$$

where e_j is the jth unit vector. Combining two inequalities above yields

$$||e_j'\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho^*)||_\infty \leq \lambda,$$

which completes the proof as the inequality holds for every j. ■

Following Bradic et al. (2021), by Assumption 8 we can define $S_{\bar{\rho}} \subset S$ as indices of a sparse approximation with $|S_{\bar{\rho}}| = \bar{s}$, where $|A|$ denotes the cardinality of set A, and coefficients $\bar{\rho} = (\bar{\rho}_1, \ldots, \bar{\rho}_p)'$, with $\bar{\rho}_j = 0$ for $j \notin S_{\bar{\rho}}$ such that

$$||\rho_L - \bar{\rho}||^2 \leq C\bar{s}\varepsilon_n^2.$$

Also define $\rho_\star$ as

$$\rho_\star = \underset{v \in \mathbb{R}^p}{\text{argmin}} (\rho_L - v)'G'\Omega_q G(\rho_L - v) + 2\varepsilon_n \sum_{j \in S_{\bar{\rho}}^c} |v_j|. \tag{B.4.1}$$

Moreover, we assume that $|\rho_\star|_1 = O(1)$.

Lemma B.4.3. $||G'\Omega_q G(\rho_\star - \rho_L)||_\infty \leq \varepsilon_n$.

Proof. Follows directly from the proof of Lemma B.4.2. ■

Lemma B.4.4. $(\rho_L - \rho_\star)'G'\Omega_q G(\rho_L - \rho_\star) \leq C\bar{s}\varepsilon_n^2$.

Proof. By the definition of $\rho_\star$ and the fact that the largest eigenvalue of $G'\Omega_q G$ is bounded, we have

$$\begin{aligned}(\rho_L - \rho_\star)'G'\Omega_q G(\rho_L - \rho_\star) + 2\varepsilon_n \sum_{j \in S_{\bar\rho}^c} |\rho_{\star,j}| &\leq (\rho_L - \bar\rho)'G'\Omega_q G(\rho_L - \bar\rho) + 2\varepsilon_n \sum_{j \in S_{\bar\rho}^c} |\bar\rho_j| \\ &= (\rho_L - \bar\rho)'G'\Omega_q G(\rho_L - \bar\rho) \\ &\leq C||\rho_L - \bar\rho||^2 \leq C\bar{s}\varepsilon_n^2.\end{aligned}$$

■

Lemma B.4.5. Let $S_{\rho_\star}$ be the vector of indices of nonzero elements of $\rho_\star$. Then, $s_\star \equiv |S_{\rho_\star}| \leq C\bar{s}$.

Proof. For all $j \in S_{\rho_\star} \backslash S_{\bar\rho}$ the first order conditions to equation (B.4.1) imply $|e_j' G'\Omega_q G(\rho_\star - \rho_L)| = \varepsilon_n$. Therefore, it follows that

$$\sum_{j \in S_{\rho_\star} \backslash S_{\bar\rho}} \left(e_j' G'\Omega_q G(\rho_\star - \rho_L)\right)^2 = \varepsilon_n^2 |S_{\rho_\star} \backslash S_{\bar\rho}|.$$

Moreover, using Lemma B.4.5 and the fact that the largest eigenvalue of $G'\Omega_q G$ is bounded, we get

$$\begin{aligned}\sum_{j \in S_{\rho_\star} \backslash S_{\bar\rho}} \left(e_j' G'\Omega_q G(\rho_\star - \rho_L)\right)^2 &\leq \sum_{j=1}^{p} \left(e_j' G'\Omega_q G(\rho_\star - \rho_L)\right)^2 \\ &= (\rho_\star - \rho_L)'G'\Omega_q G \left(\sum_{j=1}^{p} e_j e_j'\right) G'\Omega_q G(\rho_\star - \rho_L) \\ &= (\rho_\star - \rho_L)(G'\Omega_q G)^2(\rho_\star - \rho_L) \\ &\leq \lambda_{max}(G'\Omega_q G)\{(\rho_\star - \rho_L)G'\Omega_q G(\rho_\star - \rho_L)\} \leq C\bar{s}\varepsilon_n^2.\end{aligned}$$

Combining the results above, we obtain

$$\varepsilon_n^2 |S_{\rho_\star} \backslash S_{\bar\rho}| \leq C\bar{s}\varepsilon_n^2.$$

Dividing both sides by ε_n^2 gives $|S_{\rho_\star} \backslash S_{\bar\rho}| \leq C\bar{s}$. As a result,

$$s_\star = |S_{\bar\rho}| + |S_{\rho_\star} \backslash S_{\bar\rho}| \leq \bar{s} + C\bar{s} \leq C\bar{s}.$$

■

Lemma B.4.6. Let $B = \mathbb{E}[b(Z)b(Z)']$ has its largest eigenvalue bounded uniformly in n, then

$||\alpha_0 - b'\rho_\star||^2 \leq C\bar{s}\varepsilon_n^2$.

Proof. By the triangle inequality and Assumption 8,

$$\begin{aligned}||\alpha_0 - b'\rho_\star||^2 &\leq ||\alpha_0 - b'\bar{\rho}||^2 + ||b'(\bar{\rho} - \rho_L)||^2 + ||b'(\rho_L - \rho_\star)||^2 \\ &\leq C\bar{s}\varepsilon_n^2 + ||b'(\bar{\rho} - \rho_L)||^2 + ||b'(\rho_L - \rho_\star)||^2.\end{aligned}$$

Moreover, by the definition of $\bar{\rho}$ and $\lambda_{max}(B) \leq C$,

$$||b'(\bar{\rho} - \rho_L)||^2 \leq \lambda_{max}(B)||\bar{\rho} - \rho_L||^2 \leq C\bar{s}\varepsilon_n^2.$$

Also, by Lemma B.4.4,

$$||b'(\rho_L - \rho_\star)||^2 \leq \lambda_{max}(B)||\rho_L - \rho_\star||^2 \leq C\bar{s}\varepsilon_n^2,$$

which completes the proof. ■

Lemma B.4.7. If Assumptions 5–7 and 10 are satisfied, then

$$||\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho_\star)||_\infty = O_p(\varepsilon_n).$$

Proof. By the triangle inequality,

$$\begin{aligned}||\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho_\star)||_\infty &\leq ||\hat{G}'\hat{\Omega}_q\hat{M} - G'\Omega_q M||_\infty && \text{(B.4.2)}\\ &+ ||G'\Omega_q M - G'\Omega_q G\rho_\star||_\infty && \text{(B.4.3)}\\ &+ ||(G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G})\rho_\star||_\infty. && \text{(B.4.4)}\end{aligned}$$

Consider the first element (B.4.2). Note that by the triangle inequality,

$$\begin{aligned}||\hat{G}'\hat{\Omega}_q\hat{M} - G'\Omega_q M||_\infty &\leq ||(\hat{G} - G)'(\hat{\Omega}_q - \Omega_q)(\hat{M} - M)||_\infty && \text{(B.4.5)}\\ &+ ||(\hat{G} - G)'\Omega_q(\hat{M} - M)||_\infty && \text{(B.4.6)}\\ &+ ||G'(\hat{\Omega}_q - \Omega_q)(\hat{M} - M)||_\infty && \text{(B.4.7)}\\ &+ ||G'\Omega_q(\hat{M} - M)||_\infty && \text{(B.4.8)}\\ &+ ||(\hat{G} - G)'\Omega_q M||_\infty && \text{(B.4.9)}\\ &+ ||(\hat{G} - G)'(\hat{\Omega}_q - \Omega_q)M||_\infty && \text{(B.4.10)}\\ &+ ||G'(\hat{\Omega}_q - \Omega_q)M||_\infty. && \text{(B.4.11)}\end{aligned}$$

Now we will bound every term on the RHS of the inequality above. To do so, we will use

the following matrix norm inequality from Caner and Kock (2018). For any $q \times p$ matrix A, $p \times q$ matrix B, and $q \times q$ matrix F the following inequality holds

$$||BFA||_\infty \leq q||B||_\infty||F||_{\ell_\infty}||A||_\infty. \tag{B.4.12}$$

We can use (B.4.12) to put an upper bound on (B.4.5),

$$\begin{aligned}||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)(\hat{M}-M)||_\infty &\leq ||\hat{G}-G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||\hat{M}-M||_\infty \\ &= O_p(\varepsilon_n^G)o_p(1)O_p(\varepsilon_n^M) = o_p(\varepsilon_n^2).\end{aligned}$$

Moreover, notice that Assumptions 6 and 10 imply that $||G||_\infty = O(1)$ and $||M||_\infty = O(1)$. Using this fact and (B.4.12), we can bound the remaining terms (B.4.6)–(B.4.11),

$$\begin{aligned}||(\hat{G}-G)'\Omega_q(\hat{M}-M)||_\infty &\leq ||\hat{G}-G||_\infty||\Omega||_{\ell_\infty}||\hat{M}-M||_\infty = O_p(\varepsilon_n^G)O(1)O_p(\varepsilon_n^M) = O_p(\varepsilon_n^2) \\ ||G'(\hat{\Omega}_q-\Omega_q)(\hat{M}-M)||_\infty &\leq ||G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||\hat{M}-M||_\infty = O(1)o_p(1)O_p(\varepsilon_n^M) = o_p(\varepsilon_n^M) \\ ||G'\Omega_q(\hat{M}-M)||_\infty &\leq ||G||_\infty||\Omega||_{\ell_\infty}||\hat{M}-M||_\infty = O(1)O_p(\varepsilon_n^M) = O_p(\varepsilon_n^M) \\ ||(\hat{G}-G)'\Omega_q M||_\infty &\leq ||\hat{G}-G||_\infty||\Omega||_{\ell_\infty}||M||_\infty = O_p(\varepsilon_n^G)O(1) = O_p(\varepsilon_n^G) \\ ||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)M||_\infty &\leq ||\hat{G}-G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||M||_\infty = O_p(\varepsilon_n^G)o_p(1)O(1) = o_p(\varepsilon_n^G) \\ ||G'(\hat{\Omega}_q-\Omega_q)M||_\infty &\leq ||G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||M||_\infty = O(1)o_p(1) = o_p(1).\end{aligned}$$

Collecting all the terms gives the upper bound for (B.4.2)

$$||\hat{G}'\hat{\Omega}_q\hat{M} - G'\Omega_q M||_\infty = O_p(\varepsilon_n).$$

Next, by the triangle and Hölder's inequalities,

$$||G'\Omega_q M - G'\Omega_q G\rho_\star||_\infty \leq ||G'\Omega_q M - G'\Omega_q G\rho_L||_\infty + ||G'\Omega_q G(\rho_L - \rho_\star)||_\infty.$$

By Lemma B.4.2 and the fact that ρ_L are the population PGMM coefficients,

$$||G'\Omega_q M - G'\Omega_q G\rho_L||_\infty \leq \varepsilon_n.$$

Moreover, by Lemma B.4.3,

$$||G'\Omega_q G(\rho_L - \rho_\star)||_\infty \leq \varepsilon_n.$$

Thus, using the results above,

$$||G'\Omega_q M - G'\Omega_q G\rho_\star||_\infty = O(\varepsilon_n).$$

We are left with putting an upper bound on (B.4.4). By Hölder's inequality,

$$||(G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G})\rho_\star||_\infty \leq ||G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G}||_\infty|\rho_\star|_1.$$

Moreover, by the triangle inequality,

$$\begin{aligned}||G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G}||_\infty \leq\ & ||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)(\hat{G}-G)||_\infty \\ &+ 2||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)G||_\infty \\ &+ ||(\hat{G}-G)'\Omega_q(\hat{G}-G)||_\infty \\ &+ 2||(\hat{G}-G)'\Omega_q G||_\infty \\ &+ ||G'(\hat{\Omega}_q-\Omega_q)G||_\infty.\end{aligned}$$

Using (B.4.12), we can bound all the terms on the RHS of the inequality above,

$$\begin{aligned}||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)(\hat{G}-G)||_\infty &\leq ||\hat{G}-G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||\hat{G}-G||_\infty = O_p((\varepsilon_n^G)^2)o_p(1) = o_p((\varepsilon_n^G)^2) \\ 2||(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)G||_\infty &\leq 2||\hat{G}-G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||G||_\infty = O_p(\varepsilon_n^G)o_p(1)O(1) = o_p(\varepsilon_n^G) \\ ||(\hat{G}-G)'\Omega_q(\hat{G}-G)||_\infty &\leq ||\hat{G}-G||_\infty||\Omega||_{\ell_\infty}||\hat{G}-G||_\infty = O_p((\varepsilon_n^G)^2)O(1) = O_p((\varepsilon_n^G)^2) \\ 2||(\hat{G}-G)'\Omega_q G||_\infty &\leq 2||\hat{G}-G||_\infty||\Omega||_{\ell_\infty}||G||_\infty = O_p(\varepsilon_n^G)O(1) = O_p(\varepsilon_n^G) \\ ||G'(\hat{\Omega}_q-\Omega_q)G||_\infty &\leq ||G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||G||_\infty = o_p(1)O(1) = o_p(1).\end{aligned}$$

Collecting all the terms gives,

$$||(G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G})||_\infty = O_p(\varepsilon_n^G).$$

Combining the result above with $|\rho_\star|_1 = O(1)$ yields

$$||(G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G})\rho_\star||_\infty = O_p(\varepsilon_n^G)O(1) = O_p(\varepsilon_n^G).$$

Collecting all the terms for (B.4.2)–(B.4.4) gives us the desired upper bound,

$$||\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho_\star)||_\infty = O_p(\varepsilon_n) + O(\varepsilon_n) + O_p(\varepsilon_n^G) = O_p(\varepsilon_n).$$

■

Let $\phi^2(s_\star)$ denote the population restricted eigenvalue from Assumption 9 at $s = s_\star$,

$$\phi^2(s_\star) = \inf\left\{\frac{\delta' G'\Omega_q G\delta}{||\delta_{S_{\rho_\star}}||^2} : \delta \in \mathbb{R}^p\backslash\{0\},\ |\delta_{S^c_{\rho_\star}}|_1 \leq 3|\delta_{S_{\rho_\star}}|_1, |S_{\rho_\star}| \leq s_\star\right\}.$$

Next, let us introduce an empirical version of the condition above,

$$\hat{\phi}^2(s_\star) = \inf\left\{\frac{\delta'\hat{G}'\hat{\Omega}_q\hat{G}\delta}{||\delta_{S_{\rho_\star}}||^2} : \delta \in \mathbb{R}^p\backslash\{0\},\ |\delta_{S^c_{\rho_\star}}|_1 \leq 3|\delta_{S_{\rho_\star}}|_1, |S_{\rho_\star}| \leq s_\star\right\}.$$

In the following Lemma we show that we can bound $\hat{\phi}^2(s_\star)$ from below, which will be useful in the proof of Theorem 4.

Lemma B.4.8. If Assumptions 6 and 5 are satisfied, then

$$\hat{\phi}^2(s_\star) \geq \phi^2(s_\star) - O_p(s_\star \varepsilon_n^G).$$

Proof. The proof follows the proof of Lemma S3 in Caner and Kock (2018). By adding and subtracting G and Ω_q and the reverse triangle inequality,

$$\begin{aligned}
|\delta'\hat{G}'\hat{\Omega}_q\hat{G}\delta| &= |\delta'(\hat{G}-G+G)'(\hat{\Omega}_q-\Omega_q+\Omega_q)(\hat{G}-G+G)\delta| \\
&\geq |\delta' G'\Omega_q G\delta| \\
&- |\delta'(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)(\hat{G}-G)\delta| \\
&- |\delta'(\hat{G}-G)'\Omega_q(\hat{G}-G)\delta| \\
&- |\delta' G'(\hat{\Omega}_q-\Omega_q)G\delta| \\
&- 2|\delta'(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)G\delta| \\
&- 2|\delta'(\hat{G}-G)'\Omega_q G\delta|.
\end{aligned}$$

The following inequality from Caner and Kock (2018) will help us bound the expression above. For any $q \times p$ matrix A, $p \times q$ matrix B, $q \times q$ matrix F, and $p \times 1$ vector x the following inequality holds

$$|x'BFAx| \leq q|x|_1^2||B||_\infty||F||_{\ell_\infty}||A||_\infty. \tag{B.4.13}$$

Using (B.4.13), we get

$$\begin{aligned}
|\delta'(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)(\hat{G}-G)\delta| &\leq |\delta|_1^2||\hat{G}-G||_\infty^2||\hat{\Omega}-\Omega||_{\ell_\infty} \\
|\delta'(\hat{G}-G)'\Omega_q(\hat{G}-G)\delta| &\leq |\delta|_1^2||\hat{G}-G||_\infty^2||\Omega||_{\ell_\infty} \\
|\delta' G'(\hat{\Omega}_q-\Omega_q)G\delta| &\leq |\delta|_1^2||G||_\infty^2||\hat{\Omega}-\Omega||_{\ell_\infty} \\
2|\delta'(\hat{G}-G)'(\hat{\Omega}_q-\Omega_q)G\delta| &\leq 2|\delta|_1^2||\hat{G}-G||_\infty||\hat{\Omega}-\Omega||_{\ell_\infty}||G||_\infty \\
2|\delta'(\hat{G}-G)'\Omega_q G\delta| &\leq 2|\delta|_1^2||\hat{G}-G||_\infty^2||\Omega||_{\ell_\infty}||G||_\infty.
\end{aligned}$$

Combining the terms gives

$$
\begin{aligned}
|\delta'\hat{G}'\hat{\Omega}_q\hat{G}\delta| \geq\ & |\delta' G'\Omega_q G\delta| \\
& - |\delta|_1^2 ||\hat{G} - G||_\infty^2 (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty) \\
& - |\delta|_1^2 ||G||_\infty^2 ||\hat{\Omega} - \Omega||_{\ell_\infty} \\
& - 2|\delta|_1^2 ||\hat{G} - G||_\infty ||G||_\infty (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty)
\end{aligned}
\tag{B.4.14}
$$

Recall, we have the restriction

$$
|\delta_{S^c_{\rho_\star}}|_1 \leq 3|\delta_{S_{\rho_\star}}|_1 \leq 3\sqrt{s_\star}||\delta_{S_{\rho_\star}}||
$$

where the second inequality is Cauchy-Schwarz. Adding $|\delta_{S_{\rho_\star}}|$ to both sides gives

$$
|\delta|_1 \leq 4\sqrt{s_\star}||\delta_{S_{\rho_\star}}|| \Rightarrow \frac{|\delta|_1^2}{||\delta_{S_{\rho_\star}}||^2} \leq 16 s_\star. \tag{B.4.15}
$$

Divide (B.4.14) by $||\delta_{S_{\rho_\star}}||^2$ and use (B.4.15),

$$
\begin{aligned}
\frac{|\delta'\hat{G}'\hat{\Omega}_q\hat{G}\delta|}{||\delta_{S_{\rho_\star}}||^2} \geq\ & \frac{|\delta' G'\Omega_q G\delta|}{||\delta_{S_{\rho_\star}}||^2} \\
& - 16 s_\star ||\hat{G} - G||_\infty^2 (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty) \\
& - 16 s_\star ||G||_\infty^2 ||\hat{\Omega} - \Omega||_{\ell_\infty} \\
& - 32 s_\star ||\hat{G} - G||_\infty ||G||_\infty (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty).
\end{aligned}
$$

Since $\frac{|\delta' G'\Omega_q G\delta|}{||\delta_{S_{\rho_\star}}||^2} \geq \phi^2(s_\star)$ for all δ satisfying $|\delta_{S^c_{\rho_\star}}|_1 \leq 3|\delta_{S_{\rho_\star}}|_1$, minimizing the LHS of the inequality above over such δ yields

$$
\hat{\phi}^2(s_\star) \geq \phi^2(s_\star) - a_n,
$$

where

$$
\begin{aligned}
a_n =\ & 16 s_\star ||\hat{G} - G||_\infty^2 (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty) \\
& + 16 s_\star ||G||_\infty^2 ||\hat{\Omega} - \Omega||_{\ell_\infty} \\
& + 32 s_\star ||\hat{G} - G||_\infty ||G||_\infty (||\hat{\Omega} - \Omega||_{\ell_\infty} + ||\Omega||_\infty).
\end{aligned}
$$

Using Assumptions 6 and 5 and Lemma B.4.1, we can put an upper bound on a_n as follows

$$\begin{aligned}
16s_\star||\hat{G}-G||_\infty^2(||\hat{\Omega}-\Omega||_{\ell_\infty}+||\Omega||_\infty) &= 16s_\star O_p((\varepsilon_n^G)^2)(o_p(1)+O(1)) = O_p(s_\star(\varepsilon_n^G)^2)\\
16s_\star||G||_\infty^2||\hat{\Omega}-\Omega||_{\ell_\infty} &= 16s_\star O(1)o_p(1) = o_p(s_\star)\\
32s_\star||\hat{G}-G||_\infty||G||_\infty(||\hat{\Omega}-\Omega||_{\ell_\infty}+||\Omega||_\infty) &= 32s_\star O_p(\varepsilon_n^G)O(1)(o_p(1)+O(1)) = O_p(s_\star\varepsilon_n^G).
\end{aligned}$$

Gathering the terms gives

$$a_n = O_p(s_\star\varepsilon_n^G),$$

which completes the proof. ■

Proof of Theorem 4

As $\hat{\Omega}$ is positive definite, we can write

$$\hat{\rho}_L = \underset{\rho\in\mathbb{R}^q}{\operatorname{argmin}}\{||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\rho)||^2+2\lambda_n|\rho|_1\}.$$

The minimizing property of $\hat{\rho}_L$ implies

$$||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\hat{\rho}_L)||^2+2\lambda_n|\hat{\rho}_L|_1 \le ||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\rho_\star)||^2+2\lambda_n|\rho_\star|_1. \tag{B.4.16}$$

First, observe that

$$\begin{aligned}
||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\hat{\rho}_L)||^2-||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\rho_\star)||^2 &= (\hat{M}-\hat{G}\hat{\rho}_L)'\hat{\Omega}_q(\hat{M}-\hat{G}\hat{\rho}_L)-(\hat{M}-\hat{G}\rho_\star)'\hat{\Omega}_q(\hat{M}-\hat{G}\rho_\star)\\
&= \hat{\rho}_L\hat{G}'\hat{\Omega}_q\hat{G}\hat{\rho}_L-\rho_\star\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star-2(\hat{G}'\hat{\Omega}_q\hat{M})'(\hat{\rho}_L-\rho_\star)\\
&= (\hat{\rho}_L-\rho_\star)'\hat{G}'\hat{\Omega}_q\hat{G}(\hat{\rho}_L-\rho_\star)+2\rho_\star'\hat{G}'\hat{\Omega}_q\hat{G}(\hat{\rho}_L-\rho_\star)\\
&\quad -2(\hat{G}'\hat{\Omega}_q\hat{M})'(\hat{\rho}_L-\rho_\star)\\
&= ||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L-\rho_\star)||^2\\
&\quad -2(\hat{G}'\hat{\Omega}_q\hat{M}-\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star)'(\hat{\rho}_L-\rho_\star).
\end{aligned}$$

Plug the expression above in (B.4.16) to get

$$\begin{aligned}
||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L-\rho_\star)||^2+2\lambda_n|\hat{\rho}_L|_1 &\le 2(\hat{G}'\hat{\Omega}_q\hat{M}-\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star)'(\hat{\rho}_L-\rho_\star)+2\lambda_n|\rho_\star|_1\\
&\le 2||\hat{G}'\hat{\Omega}_q\hat{M}-\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star||_\infty|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1\\
&= 2||\hat{G}'\hat{\Omega}_q(\hat{M}-\hat{G}\rho_\star)||_\infty|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1\\
&= 2o_p(\lambda_n)|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1,
\end{aligned} \tag{B.4.17}$$

where the second inequality is Hölder and the last equality comes from Lemma B.4.7 and the fact

that $\varepsilon_n = o(\lambda_n)$. Hence, with probability approaching one,

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + 2\lambda_n|\hat{\rho}_L|_1 \leq 2\lambda_n|\hat{\rho}_L - \rho_\star|_1 + 2\lambda_n|\rho_\star|_1.$$

Next, note that $|\hat{\rho}_L|_1 = |\hat{\rho}_{L,S_{\rho_\star}}|_1 + |\hat{\rho}_{L,S^c_{\rho_\star}}|_1$ and $|\rho_\star|_1 = |\rho_{\star,S_{\rho_\star}}|_1$ as $|\rho_{\star,S^c_{\rho_\star}}|_1 = 0$. Therefore,

$$\begin{aligned}||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + 2\lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 &\leq 2\lambda_n|\hat{\rho}_L - \rho_\star|_1 + 2\lambda_n(|\rho_{\star,S_{\rho_\star}}|_1 - |\hat{\rho}_{L,S_{\rho_\star}}|_1) \\ &\leq 2\lambda_n|\hat{\rho}_L - \rho_\star|_1 + 2\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1,\end{aligned}$$

where the second line comes from the reverse triangle inequality. Using that $|\hat{\rho}_L - \rho_\star|_1 = |\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1 + |\hat{\rho}_{L,S^c_{\rho_\star}}|_1$ gives

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + \lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq 3\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1. \tag{B.4.18}$$

The inequality in (B.4.18) implies $\lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq 3\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1$ leading to $|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq 3|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1$, meaning that the restricted eigenvalue condition is satisfied. Note that by Cauchy-Schwarz inequality, $|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1 \leq \sqrt{s_\star}||\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}||$. Using this along with the restricted eigenvalue condition on (B.4.18) yields

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + \lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq 3\lambda_n\sqrt{s_\star}||\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|| \leq 3\lambda_n\sqrt{s_\star}\frac{||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||}{\hat{\phi}(s_\star)}.$$

Note that by AM-GM inequality,

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + \lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq \frac{1}{2}||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + \frac{9}{2}\frac{\lambda_n^2 s_\star}{\hat{\phi}^2(s_\star)}.$$

Multiplying both sides by 2 and collecting terms gives

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + 2\lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq \frac{9\lambda_n^2 s_\star}{\hat{\phi}^2(s_\star)}. \tag{B.4.19}$$

To get the ℓ_1-error bound, ignore the first term on the LHS of (B.4.18) and add $\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1$ to both sides,

$$\lambda_n|\hat{\rho}_L - \rho_\star|_1 \leq 4\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1.$$

By Cauchy-Schwarz inequality and the restricted eigenvalue condition,

$$\lambda_n|\hat{\rho}_L - \rho_\star|_1 \leq 4\lambda_n\sqrt{s_\star}||\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|| \leq 4\lambda_n\sqrt{s_\star}\frac{||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||}{\hat{\phi}(s_\star)}.$$

The bound on $||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2$ in (B.4.19) implies

$$|\hat{\rho}_L - \rho_\star|_1 \leq \frac{12\lambda_n s_\star}{\hat{\phi}^2(s_\star)}.$$

Next, by Lemma B.4.8 and $\varepsilon_n = o(\lambda_n)$,

$$|\hat{\rho}_L - \rho_\star|_1 \leq \frac{12\lambda_n s_\star}{\phi^2(s_\star) - o_p(s_\star \lambda_n)}.$$

Focus on the RHS of the inequality,

$$\frac{C\lambda_n s_\star}{\phi^2(s_\star) - o_p(s_\star \lambda_n)} = \frac{C}{\phi^2(s_\star)/(\lambda_n s_\star) - o_p(1)},$$

meaning that with probability approaching one,

$$|\hat{\rho}_L - \rho_\star|_1 \leq \frac{C\lambda_n s_\star}{\phi^2(s_\star)}.$$

Moreover, applying the result of Lemma B.4.5 gives

$$|\hat{\rho}_L - \rho_\star|_1 = O_p(\bar{s}\lambda_n). \tag{B.4.20}$$

Finally, let $\alpha_\star = b(Z)'\rho_\star$, then by the triangle inequality and Lemma B.4.6,

$$||\hat{\alpha}_L - \alpha_0||^2 \leq ||\hat{\alpha}_L - \alpha_\star||^2 + ||\alpha_\star - \alpha_0||^2 \leq ||\hat{\alpha}_L - \alpha_\star||^2 + C\bar{s}\varepsilon_n^2.$$

By Hölder's inequality and (B.4.20),

$$||\hat{\alpha}_L - \alpha_\star||^2 = (\hat{\rho}_L - \rho_\star)'B(\hat{\rho}_L - \rho_\star) \leq ||B||_\infty |\hat{\rho}_L - \rho_\star|_1^2 \leq O_p(\bar{s}^2\lambda_n^2).$$

The conclusion comes from the fact that $\bar{s}^2\lambda_n^2 > \bar{s}^2\varepsilon_n^2 \geq \bar{s}\varepsilon_n^2$, where the second inequality is due to $\bar{s}^2$ growing faster than $\bar{s}$.

■

B.4.2. Asymptotic properties

Lemma B.4.9. If Assumptions 5–7 and 10 are satisfied and $\varepsilon_n = o(\lambda_n)$, then

$$|\hat{\rho}_L|_1 = O_p(1).$$

Proof. Recall Equation (B.4.17) from the proof of Theorem 4 which implies

$$2\lambda_n|\hat{\rho}_L|_1 \leq 2o_p(\lambda_n)|\hat{\rho}_L - \rho_\star|_1 + 2\lambda_n|\rho_\star|_1.$$

Dividing both sides by $2\lambda_n$ and applying the triangle inequality gives

$$|\hat{\rho}_L|_1 \leq o_p(1)|\hat{\rho}_L - \rho_\star|_1 + |\rho_\star|_1 \leq |\rho_\star|_1 + o_p(1)(|\hat{\rho}_L|_1 + |\rho_\star|_1),$$

which implies that with probability approaching one,

$$|\hat{\rho}_L|_1 \leq |\rho_\star|_1 + \frac{1}{2}(|\hat{\rho}_L|_1 + |\rho_\star|_1).$$

Subtracting $|\hat{\rho}_L|_1/2$ from both sides and multiplying by 2 gives with probability approaching one

$$|\hat{\rho}_L|_1 \leq 3|\rho_\star|_1 = O(1).$$

■

Proof of Theorem 5

We prove the first conclusion by verifying the conditions of Lemma 15 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020). Let $g(w, \gamma, \alpha, \theta)$ and $\phi(w, \gamma, \alpha, \theta)$ in Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) be $m(w, \gamma) - \theta$ and $\alpha(z)[y - \gamma(x)]$ here, respectively. First, $\mathbb{E}[\psi(W_i, \gamma_0, \alpha_0, \theta_0)^2] < \infty$ follows from Assumption 11. Moreover, note that by Assumptions 11 and 12, Theorem 4, and the law of iterated expectations,

$$\begin{aligned}
\int [\phi(w, \hat{\gamma}_\ell, \alpha_0) - \phi(w, \gamma_0, \alpha_0)]^2 F_0(dw) &= \int \alpha_0^2(z)[\hat{\gamma}_\ell(x) - \gamma_0(x)]^2 F_0(dw) \leq C||T(\hat{\gamma}_\ell - \gamma_0)||^2 \xrightarrow{p} 0 \\
\int [\phi(w, \gamma_0, \hat{\alpha}_\ell) - \phi(w, \gamma_0, \alpha_0)]^2 F_0(dw) &= \int [\hat{\alpha}_\ell(z) - \alpha_0(z)]^2 [y - \gamma_0(x)]^2 F_0(dw) \\
&= \int [\hat{\alpha}_\ell(z) - \alpha_0(z)]^2 \mathbb{E}[[y - \gamma_0(x)]^2 | z] F_0(dz) \\
&\leq C||\hat{\alpha}_\ell - \alpha_0||^2 \xrightarrow{p} 0.
\end{aligned}$$

Also, it follows from Assumption 12 that

$$\int [m(w, \hat{\gamma}_\ell) - m(w, \gamma_0)]^2 F_0(dw) \xrightarrow{p} 0.$$

Thus, Assumption 1 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) is satisfied.

Next, for each ℓ let

$$\hat{\Delta}_\ell(w) = \phi(w, \hat{\gamma}_\ell, \hat{\alpha}_\ell) - \phi(w, \gamma_0, \hat{\alpha}_\ell) - \phi(w, \hat{\gamma}_\ell \alpha_0) + \phi(w, \gamma_0, \alpha_0) = [\hat{\alpha}_\ell(z) - \alpha_0(z)][\hat{\gamma}_\ell(x) - \gamma_0(x)].$$

Since α_0 is bounded by Assumption 11 and $\sup_z |\hat{\alpha}_\ell(z)| = O_p(1)$ by Lemma B.4.9,

$$\begin{aligned}
\int \hat{\Delta}_\ell^2(w) F_0(dw) &= \int [\hat{\alpha}_\ell(z) - \alpha_0(z)]^2 [\hat{\gamma}_\ell(x) - \gamma_0(x)]^2 F_0(dw) \\
&\leq O_p(1) \int [\hat{\gamma}_\ell(x) - \gamma_0(x)]^2 F_0(dw) \xrightarrow{p} 0,
\end{aligned}$$

where the conclusion follows from Assumption 12. Furthermore, by Cauchy-Schwarz inequality

and Assumption 14,

$$\begin{aligned}\left|\sqrt{n}\int \hat{\Delta}_\ell(w)F_0(dw)\right| &= \sqrt{n}\left|\int [\hat{\alpha}_\ell(z)-\alpha_0(z)][\hat{\gamma}_\ell(x)-\gamma_0(x)]F_0(dw)\right| \\ &= \sqrt{n}\left|\int [\hat{\alpha}_\ell(z)-\alpha_0(z)]\mathbb{E}[\hat{\gamma}_\ell(x)-\gamma_0(x)|z]F_0(dz)\right| \\ &\leq \sqrt{n}||\hat{\alpha}_\ell-\alpha_0||\ ||T(\hat{\gamma}_\ell-\gamma_0)|| = O_p(n^{1/2}\kappa_n^\alpha\kappa_n^\gamma) \xrightarrow{p} 0,\end{aligned}$$

which renders Assumption 2(iii) of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) satisfied.

Also, by construction,

$$\int \hat{\alpha}_\ell(z)[y-\gamma_0(x)]F_0(dw) = \mathbb{E}[\hat{\alpha}_\ell(z)\mathbb{E}[y-\gamma_0(x)]|z] = 0.$$

Since $m(w,\gamma)$ is affine in γ, it verifies Assumption 3 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) is satisfied. As a result, we get the first conclusion.

To get the second conclusion, we need to show that $\hat{V}$ is a consistent estimator of V. This part of the proof is very similar to the proof of Theorem 5 in Chernozhukov, Newey and Robins (2020). We start with

$$\hat{V} = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_i^2 = \frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)^2 + \frac{2}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)\psi_i + \frac{1}{n}\sum_{i=1}^n\psi_i^2,$$

hence, by re-arranging the terms and Cauchy-Schwarz inequality,

$$\hat{V}-V = \frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)^2 + \frac{2}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)\psi_i \leq \frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i-\psi_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\psi_i^2}. \tag{B.4.21}$$

Using the triangle inequality, for $i \in I_\ell$,

$$(\hat{\psi}_i-\psi_i)^2 \leq C\sum_{j=1}^4 R_{ij} = C\sum_{j=1}^3 R_{ij} + o_p(1),$$

where

$$\begin{aligned}R_{i1} &= [m(W_i,\hat{\gamma}_\ell)-m(W_i,\gamma_0)]^2, \\ R_{i2} &= \hat{\alpha}_\ell^2(Z_i)[\hat{\gamma}_\ell(X_i)-\gamma_0(X_i)]^2, \\ R_{i3} &= [\hat{\alpha}_\ell(Z_i)-\alpha_0(Z_i)]^2[Y_i-\gamma_0(X_i)]^2, \\ R_{i4} &= (\hat{\theta}-\theta_0)^2.\end{aligned}$$

The first conclusion implies $R_{i4} \xrightarrow{p} 0$. Let $I_{-\ell}$ denote observations not in I_ℓ.

By Markov's inequality, for some $\delta > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_i)^2 > \delta\right) \leq \frac{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_i)^2\right]}{\delta}.$$

Note that the cross-fitting allows us to write

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_i)^2\right] \leq \mathbb{E}\left[\frac{C}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\sum_{j=1}^{3}R_{ij}\right] + o_p(1) = C\sum_{\ell=1}^{L}\frac{n_\ell}{n}\sum_{j=1}^{3}\mathbb{E}[\mathbb{E}[R_{ij}|I_{-\ell}]] + o_p(1).$$

Furthermore, by Hölder's inequality and Assumption 6,

$$\max_{i\in I_\ell}|\hat{\alpha}_\ell(Z_i)| \leq |\hat{\rho}_L|_1 \max_{i\in I_\ell}||b(Z_i)||_\infty \leq C_b|\hat{\rho}_L|_1.$$

By Lemma B.4.9,

$$\max_i |\hat{\alpha}_\ell(Z_i)| = C_b O_p(\bar{A}_n) = O_p(1).$$

Then for $i \in I_\ell$ by Assumptions 11, 12, and iterated expectations,

$$\begin{aligned}
\mathbb{E}[R_{i1}|I_{-\ell}] &= \int [m(W_i, \hat{\gamma}_\ell) - m(W_i, \gamma_0)]^2 F_0(dW) \xrightarrow{p} 0, \\
\mathbb{E}[R_{i2}|I_{-\ell}] &\leq O_p(1)\int [\hat{\gamma}_\ell(X_i) - \gamma_0(X_i)]^2 F_0(dX) \xrightarrow{p} 0, \\
\mathbb{E}[R_{i3}|I_{-\ell}] &= \mathbb{E}\left[\mathbb{E}\left[[\hat{\alpha}_\ell(Z_i) - \alpha_0(Z_i)]^2[Y_i - \gamma_0(X_i)]^2|Z_i, I_{-\ell}\right]|I_{-\ell}\right] \\
&= \mathbb{E}\left[[\hat{\alpha}_\ell(Z_i) - \alpha_0(Z_i)]^2\mathbb{E}[[Y_i - \gamma_0(X_i)]^2|Z_i]|I_{-\ell}\right] \\
&\leq C||\hat{\alpha}_\ell - \alpha_0||^2 \xrightarrow{p} 0.
\end{aligned}$$

As a result,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\psi}_i - \psi_i)^2 \xrightarrow{p} 0.$$

Furthermore, $\mathbb{E}[\psi_i^2] < \infty$ by Assumptions 11 and 12. Thus, the conclusion follows from (B.4.21) and the central limit theorem.

■

Proof of Lemma 4

The proof is similar to the proof of Lemma 10 of Chernozhukov, Newey and Robins (2020). We start with defining

$$\hat{M}_\ell = (\hat{M}_{\ell 1}, \ldots, \hat{M}_{\ell q})', \; \hat{M}_{\ell j} = \frac{1}{n - n_\ell}\sum_{\ell' \neq \ell}\sum_{i\in I_{\ell'}} D(W_i, d_j, \tilde{\gamma}_{\ell,\ell'}),$$

$$\bar{M}_\ell(\gamma) = (\bar{M}_{\ell 1}(\gamma), \ldots, \bar{M}_{\ell q}(\gamma))', \; \bar{M}_{\ell j} = \int D(w, d_j, \gamma)F_0(dw).$$

Note that $M = \bar{M}(\gamma_0)$. Let $\Gamma_{\ell,\ell'} = \{||\tilde{\gamma}_{\ell,\ell'} - \gamma_0|| \leq \varepsilon\}$, and note that $\mathbb{P}(\Gamma_{\ell,\ell'}) \to 1$ for each ℓ and ℓ' by Assumption 15. When $\Gamma_{\ell,\ell'}$ occurs,

$$\max_{1\leq j\leq q} |D(W, d_j, \gamma)| \leq C$$

by Assumption 15. For $i \in I_{\ell'}$ define

$$T_{ij}(\gamma) = D(W_i, d_j, \gamma) - \bar{M}(\gamma), U_{ij}(\gamma) = \frac{1}{n_{\ell'}} \sum_{i\in I_{\ell'}} T_{ij}(\gamma).$$

Note that for any constant $\bar{C}$ and the event $\mathcal{A} = \{\max_{1\leq j\leq q} |U_{ij}(\gamma)| \geq \bar{C}\varepsilon_n\}$ where $\varepsilon_n = \sqrt{\log(q)/n}$,

$$\begin{aligned}\mathbb{P}(\mathcal{A}) &= \mathbb{P}(\mathcal{A}|\Gamma_{\ell,\ell'})\mathbb{P}(\Gamma_{\ell,\ell'}) + \mathbb{P}(\mathcal{A}|\Gamma^c_{\ell,\ell'})[1 - \mathbb{P}(\Gamma_{\ell,\ell'})] \\ &\leq \mathbb{P}\left(\max_{1\leq j\leq q} |U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n|\Gamma_{\ell,\ell'}\right) + [1 - \mathbb{P}(\Gamma_{\ell,\ell'})].\end{aligned} \tag{B.4.22}$$

Moreover,

$$\mathbb{P}\left(\max_{1\leq j\leq q} |U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n|\Gamma_{\ell,\ell'}\right) \leq q \max_{1\leq j\leq q} \mathbb{P}\left(|U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n|\Gamma_{\ell,\ell'}\right).$$

Note that $\mathbb{E}[T_{ij}(\tilde{\gamma}_{\ell,\ell'})|\tilde{\gamma}_{\ell,\ell'}] = 0$ for $i \in I_{\ell'}$. Furthermore, conditional on $\Gamma_{\ell,\ell'}$, for $i \in I_{\ell'}$,

$$|T_{ij}(\tilde{\gamma}_{\ell,\ell'})| \leq 2C.$$

Hence, T_{ij} is bounded. Similar to the proof of Lemma B.4.1, define $K = 2C/\log 2 \geq ||T_{ij}||_{\Psi_2}$. By Hoeffding's inequality (see Theorem 2.6.2 in Vershynin (2018)) and the independence of $\{W_i\}_{i\in I_{\ell'}}$ and $\tilde{\gamma}_{\ell,\ell'}$, there is a constant c such that

$$\begin{aligned}q \max_{1\leq j\leq q} \mathbb{P}\left(|U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n|\Gamma_{\ell,\ell'}\right) &= q\mathbb{E}\left[\max_{1\leq j\leq q} \mathbb{P}\left(|U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n|\tilde{\gamma}_{\ell,\ell'}\right)\middle|\Gamma_{\ell,\ell'}\right] \\ &\leq 2q\mathbb{E}\left[\exp\left(-\frac{c(n_{\ell'}\bar{C}\varepsilon_n)^2}{n_{\ell'}K^2}\right)\middle|\Gamma_{\ell,\ell'}\right] \\ &\leq 2q\exp\left(-\frac{cn_{\ell'}\bar{C}^2\log(q)}{Ln_{\ell'}K^2}\right) \\ &\leq 2\exp\left(\log(q)\left[1 - \frac{c\bar{C}^2}{LK^2}\right]\right) \to 0\end{aligned}$$

for any $\bar{C} > K\sqrt{L/c}$. Let $U_{\ell'}(\gamma) = (U_{\ell'1}, \ldots, U_{\ell'q})'$. Then it follows from (B.4.22) that for large $\bar{C}$, $\mathbb{P}(|U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n) \to 0$, meaning that $||U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})||_\infty = O_p(\varepsilon_n)$.

Next, for each ℓ by the triangle inequality we have,

$$||\hat{M}_\ell - M||_\infty \leq ||\hat{M}_\ell - \bar{M}(\tilde{\gamma}_{\ell,\ell'})||_\infty + ||\bar{M}(\tilde{\gamma}_{\ell,\ell'}) - M||_\infty.$$

Furthermore, $n - n_\ell = \sum_{\ell' \neq \ell} n_{\ell'}$ and

$$||\hat{M}_\ell - \bar{M}(\tilde{\gamma}_{\ell,\ell'})||_\infty = \left\| \hat{M}_\ell - \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_\ell} \bar{M}(\tilde{\gamma}_{\ell,\ell'}) \right\|_\infty \leq \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_\ell} ||U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})||_\infty = O_p(\varepsilon_n).$$

Also, by Assumption 15(ii) and $\mathbb{P}(\Gamma_{\ell,\ell'}) \to 1$ for each ℓ and ℓ',

$$||\bar{M}(\tilde{\gamma}_{\ell,\ell'}) - M||_\infty \leq \left\| \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_\ell} [\bar{M}(\tilde{\gamma}_{\ell,\ell'}) - M] \right\|_\infty \leq C \sum_{\ell' \neq \ell} \frac{n_{\ell'}}{n - n_\ell} ||\tilde{\gamma}_{\ell,\ell'} - \gamma_0|| = O_p(\kappa_n^\gamma).$$

The conclusion follows from κ_n^γ being a slower rate than ε_n. ■

Proof of Theorem 6

The proof is analogous to the proof of Theorem 5. We obtain the first conclusion by verifying the conditions of Lemma 15 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020). First, it follows from the proof of Theorem 5 that the conditions of Assumptions 1 and 2 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) are satisfied.

Next, by Assumptions 16 and 17,

$$\begin{aligned}
\sqrt{n}|\bar{\psi}(w, \hat{\gamma}_\ell, \alpha_0, \theta_0)| &= \sqrt{n} \left| \int [m(w, \hat{\gamma}_\ell) - \theta_0 + \alpha_0(z)[y - \hat{\gamma}_\ell(x)]] F_0(dw) \right| \\
&= \sqrt{n} \left| \int [m(w, \hat{\gamma}_\ell) - m(w, \gamma_0) + \alpha_0(z)[y - \hat{\gamma}_\ell(x)]] F_0(dw) \right| \\
&= \sqrt{n} \left| \int [m(w, \hat{\gamma}_\ell) - m(w, \gamma_0) + \alpha_0(z)[\gamma_0(x) - \hat{\gamma}_\ell(x)]] F_0(dw) \right| \\
&= \sqrt{n} \left| \int [m(w, \hat{\gamma}_\ell) - m(w, \gamma_0) - D(w, \gamma_0, \hat{\gamma}_\ell - \gamma_0)] F_0(dw) \right| \\
&\leq C\sqrt{n} ||\hat{\gamma}_\ell - \gamma_0||^2 \\
&= \sqrt{n} o_p((n^{-1/4})^2) = o_p(1).
\end{aligned}$$

Moreover, as in the proof of Theorem 5,

$$\int \hat{\alpha}_\ell(z)[y - \gamma_0(x)] F_0(dw) = 0.$$

Thus, Assumption 3 of Chernozhukov, Escanciano, Ichimura, Newey and Robins (2020) is satisfied, which combined with the results above gives us the first conclusion. The second conclusion follows exactly as in the proof of Theorem 5. ■

APPENDIX C: Additional Details for Chapter 3

C.1. Data cleaning and aggregation details

C.1.1. Imputations

Data on product characteristics have a lot of missing observations in the type of sweetener and caffeine level. We use the following heuristics to impute those values:

- TYPE OF SWEETENER:
  - if the calorie level is "REGULAR", then the type of sweetener will be "SUGAR";
  - if the calorie level is "CALORIE-FREE", then the type of sweetener will be " UNSWEETENED";
  - if the calorie level is diet and the flavor is not cola, then the type of sweetener will be "SWEETENER".
- CAFFEINE INFO:
  - if flavor is "GRAPEFRUIT", "LEMON LIME", "NATURAL", "STRAWBERRY", "PINEAPPLE", "GRAPE", "FRUIT PUNCH", it is "CAFFEINE FREE";
  - if flavor is "DEW" , "PEPPER", "CHERRY COLA", it is "CAFFEINE".

We also replace zero sales with ones and impute corresponding missing prices with the average price of all other observed products in a particular store in a particular week.

C.1.2. Product characteristics aggregation

All product characteristics are categorical variables, to facilitate computations we group product attributes into larger groups which can be coded up as dummy variables. We use the following heuristics:

- FLAVOR/SCENT:
  - cola (such as "CHERRY COLA", "COLA WITH LEMON" and so on, basically everything with "COLA")
  - lemonade (such as "LEMONADE", "LEMON LIME", "MANDARINE LIME", "CITRUS", "TANGERINE", "PUNCH", etc.)

  - alcohol-free beer (such as "ROOT BEER","BIRCH BEER", etc.)
  - berries ("STRAWBERRY", "RASPBERRY" , "CHERRY", etc.)
  - fruit (fruity flavors except berries or lemon, such as "PINEAPPLE", "GRAPE", "PEACH", "WATERMELON", etc.)
  - cream soda ("RED CREAM SODA", "CREAM SODA", etc.)
  - others
- CALORIE LEVEL:
  - caffeine free and 55% caffeine free are considered caffeine free
  - other beverages are considered to contain caffeine
- CAFFEINE LEVEL:
  - calorie free and diet beverages are considered to be diet
  - other beverages are considered to be regular
- TYPE OF SWEETENER:
  - sugar free
  - sweetener (non-saccharin): Nutra, aspartame, sucralose, splenda
  - sugar and/or corn sweetener/syrup: contains all entries corresponding to corn sweeteners and sugar/saccharin containing products

C.2. GNT basis functions

Here we present an idea behind the approximation strategy in GNT. We have a function $\gamma(\omega_{jt})$ we need to approximate, where

$$\omega_{jt} = \left(\omega'_{j,1,t}, \ldots, \omega'_{j,j-1,t}, \omega'_{j,j+1,t}, \ldots, \omega'_{j,J,t}\right)'$$

is a vector representing the "state" of product j in market t (the shares and product characteristic differences with respect to the rivals in the same market). Given the vector symmet-

ric theory underlying demand across markets, without loss of generality we can express

$$\gamma\left(\omega_{jt}\right) = g\left(F\left(\omega_{jt}\right)\right)$$

where F is the empirical distribution of the variables in ω_{jt}.

An approximation strategy for γ can be structured as following. For simplicity, write $F_{jt} = F\left(\omega_{jt}\right)$ and let us approximate the distribution F_{jt} by a finite set of moments $m_1\left(F_{jt}\right), \ldots,$ $m_L\left(F_{jt}\right)$. Then our approximation to γ can be expressed as

$$\gamma\left(\omega_{jt}\right) \approx g\left(m_1\left(F_{jt}\right), \ldots, m_L\left(F_{jt}\right)\right).$$

There are two issues we need to resolve to implement this approximation:

1. The choice of moments $m_1\left(F_{jt}\right), \ldots, m_L\left(F_{jt}\right)$
2. The choice of a predictive function g

Let us first deal with the choice of m_l, $l = 1, \ldots, L$. Let us define $M_{jt}\left(\tau\right)$ as the MGF associated with F_{jt}, where $\tau = \left(\tau_1, \ldots, \tau_{d_{x_2}+1}\right)$ and d_{x_2} is the dimension of $x^{(2)}$. Then define the moment

$$m^{jt}_{p_1,\ldots,p_{d_{x_2}+1}} = \left.\frac{\partial^{p_1+\ldots p_{d_{x_2}+1}}}{\partial t_1^{p_1}\ldots\partial t_{d_{x_2}+1}^{p_{d_{x_2}+1}}} M_{jt}(\tau)\right|_{\tau=0}.$$

This class of moments is defined by the multi-index $p_1, \ldots p_{d_{x_2}+1}$ for $p_k \in \mathbb{Z}_+$. We can define the set of nth order moments to be

$$B^{jt}_n = \left\{ m^{jt}_{p_1,\ldots,p_{d_{x_2}+1}} : \sum_{k=1}^{d_{x_2}+1} p_k = n \text{ and } n \geq 2 \text{ and } p_1 > 0 \text{ and } \exists k > 1 \; s.t. \; p_k > 0 \right\}.$$

Observe that we restrict shares which are the first dimension of the state vector ω_{jt} to never enter with a zero power, e.g., each moment has some interaction with shares. In addition, shares must interact with at least one dimension of differentiation. Then the set of moments entering the nth order approximation for each t is

$$\bigcup_{i=2}^{n} B^{jt}_i$$

The choice of g can be determine by any functional form that allows for a flexible ap-

proximation from the predictors $m_1(F_{jt}), \ldots, m_L(F_{jt})$, such as polynomials, B-splines, wavelets, etc.

We use the idea above to construct $b(z_{jt})$ and $d(\omega_{jt})$ dictionaries. We use 3rd order moments to construct $b(z_{jt})$ and 2nd order moments to construct $d(\omega_{jt})$. Then we construct quadratic polynomials with interaction terms based on these moments, which gives $p = 405$ and $q = 594$.

BIBLIOGRAPHY

Ackerberg, D., Chen, X., Hahn, J. and Liao, Z. (2014), 'Asymptotic efficiency of semiparametric two-step gmm', *Review of Economic Studies* **81**(3), 919–943.

Ai, C. and Chen, X. (2003), 'Efficient estimation of models with conditional moment restrictions containing unknown functions', *Econometrica* **71**(6), 1795–1843.

Ai, C. and Chen, X. (2007), 'Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables', *Journal of Econometrics* **141**(1), 5–43.

Bakhitov, E., Gandhi, A. and Tao, J. (2020), Feature selection in differentiated product demand models, Technical report, Working paper.

Bakhitov, E. and Singh, A. (2021), 'Causal gradient boosting: Boosted instrumental variable regression', *arXiv preprint arXiv:2101.06078* .

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**(6), 2369–2429.

Bennett, A., Kallus, N. and Schnabel, T. (2019), Deep generalized method of moments for instrumental variable analysis, *in* 'Advances in Neural Information Processing Systems', pp. 3559–3569.

Berry, S. (1994), 'Estimating discrete-choice models of product differentiation', *The RAND Journal of Economics* pp. 242–262.

Berry, S., Gandhi, A. and Haile, P. (2013), 'Connected substitutes and invertibility of demand', *Econometrica* **81**(5), 2087–2111.

Berry, S. and Haile, P. (2016), 'Identification in differentiated products markets', *Annual review of Economics* **8**, 27–52.

Berry, S., Levinsohn, J. and Pakes, A. (1995), 'Automobile prices in market equilibrium', *Econometrica: Journal of the Econometric Society* pp. 841–890.

Berry, S. T. and Haile, P. A. (2014), 'Identification in differentiated products markets using market level data', *Econometrica* **82**(5), 1749–1797.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993), *Efficient and adaptive estimation for semiparametric models*, Vol. 4, Johns Hopkins University Press Baltimore.

Bickel, P. J., Ritov, Y., Tsybakov, A. B. et al. (2009), 'Simultaneous analysis of lasso and dantzig selector', *The Annals of statistics* **37**(4), 1705–1732.

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Semi-nonparametric iv estimation of shape-invariant engel curves', *Econometrica* **75**(6), 1613–1669.

Bradic, J., Chernozhukov, V., Newey, W. K. and Zhu, Y. (2021), 'Minimax semiparametric learning with approximate sparsity', *arXiv preprint arXiv:1912.12213* .

Breiman, L. (1998), 'Arcing classifiers', *The annals of statistics* **26**(3), 801–849.

Breiman, L. (1999), 'Prediction games and arcing algorithms', *Neural computation* **11**(7), 1493–1517.

Bronnenberg, B. J., Kruger, M. W. and Mela, C. F. (2008), 'Database paper—the iri marketing data set', *Marketing science* **27**(4), 745–748.

Bühlmann, P. and Hothorn, T. (2007), 'Boosting algorithms: Regularization, prediction and model fitting', *Statistical Science* **22**(4), 477–505.

Bühlmann, P. and Yu, B. (2003), 'Boosting with the l2 loss: regression and classification', *Journal of the American Statistical Association* **98**(462), 324–339.

Caner, M. and Kock, A. B. (2018), 'High dimensional linear gmm', *arXiv preprint arXiv:1811.08779* .

Carrasco, M., Florens, J.-P. and Renault, E. (2007), 'Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization', *Handbook of econometrics* **6**, 5633–5751.

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics* **34**(3), 305–334.

Chen, J., Chen, X. and Tamer, E. (2021), 'Efficient estimation in npiv models: A comparison of various neural networks-based estimators', *arXiv preprint arXiv:2110.06763* .

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', pp. 785–794.

Chen, X. and Christensen, T. M. (2018), 'Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression', *Quantitative Economics* **9**(1), 39–84.

Chen, X. and Pouzo, D. (2012), 'Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals', *Econometrica* **80**(1), 277–321.
URL: *http://dx.doi.org/10.3982/ECTA7888*

Chen, X. and Pouzo, D. (2015), 'Sieve wald and qlr inferences on semi/nonparametric conditional moment models', *Econometrica* **83**(3), 1013–1079.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2020), 'Locally robust semiparametric estimation', *arXiv preprint arXiv:1608.00033* .

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V. and Syrgkanis, V. (2021), 'Automatic debiased machine learning via neural nets for generalized linear regression', *arXiv preprint arXiv:2104.14737* .

Chernozhukov, V., Newey, W. K. and Robins, J. (2018), Double/de-biased machine learning using regularized riesz representers, Technical report, cemmap working paper.

Chernozhukov, V., Newey, W. K. and Robins, J. (2020), Automatic debiased machine learning of causal and structural effects, Technical report.

Chernozhukov, V., Newey, W. K., Robins, J. and Singh, R. (2019), 'Double/de-biased machine learning of global and local parameters using regularized riesz representers', *stat* **1050**, 9.

Chernozhukov, V., Newey, W. K. and Singh, R. (2018), 'Learning l2 continuous regression functionals via regularized riesz representers', *arXiv preprint arXiv:1809.05224* .

Chernozhukov, V., Newey, W., Singh, R. and Syrgkanis, V. (2020), 'Adversarial estimation of riesz representers', *arXiv preprint arXiv:2101.00009* .

Compiani, G. (2018), 'Nonparametric demand estimation in differentiated products markets', *Available at SSRN 3134152* .

Conlon, C. and Gortmaker, J. (2020), 'Best practices for differentiated products demand estimation with pyblp', *The RAND Journal of Economics* **51**(4), 1108–1161.

Darolles, S., Fan, Y., Florens, J.-P. and Renault, E. (2011), 'Nonparametric instrumental regression', *Econometrica* **79**(5), 1541–1565.

Dikkala, N., Lewis, G., Mackey, L. and Syrgkanis, V. (2020), 'Minimax estimation of conditional moment models', *arXiv preprint arXiv:2006.07201* .

Fosgerau, M., Monardo, J. and De Palma, A. (2020), 'The inverse product differentiation logit model', *Available at SSRN 3141041* .

Freund, Y. (1995), 'Boosting a weak learning algorithm by majority', *Information and computation* **121**(2), 256–285.

Freund, Y. and Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* 'icml', Vol. 96, Citeseer, pp. 148–156.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences* **55**(1), 119–139.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Friedman, J. H. and Popescu, B. (2003), 'Importance sampled learning ensembles', *Journal of Machine Learning Research* **94305**, 1–32.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007), 'Pathwise coordinate optimization', *The annals of applied statistics* **1**(2), 302–332.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2000), 'Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)', *The annals of statistics* **28**(2), 337–407.

Gandhi, A., Hosanagar, K. and Singh, A. (2019), 'Learning optimal instrument variables', *Available at SSRN 3352957* .

Gandhi, A. and Houde, J.-F. (2019), Measuring substitution patterns in differentiated products industries, Technical report, National Bureau of Economic Research.

Gandhi, A., Nevo, A. and Tao, J. (2020), Flexible estimation of differentiated product demand models using aggregate data, Technical report, Working paper.

Gautier, E. and Rose, C. (2021), 'High-dimensional instrumental variables regression and confidence sets', *arXiv preprint arXiv:1105.2454* .

Gold, D., Lederer, J. and Tao, J. (2020), 'Inference for high-dimensional instrumental variables regression', *Journal of Econometrics* **217**(1), 79–111.

Hall, P. and Horowitz, J. L. (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *The Annals of Statistics* **33**(6), 2904–2929.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017*a*), Deep iv: A flexible approach for counterfactual prediction, *in* 'Proceedings of the 34th International Conference on Machine Learning-Volume 70', JMLR. org, pp. 1414–1423.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017*b*), Deep iv: A flexible approach for counterfactual prediction, *in* 'Proceedings of the 34th International Conference on Machine Learning-Volume 70', JMLR. org, pp. 1414–1423.

Hausman, J. A. and Newey, W. K. (1995), 'Nonparametric estimation of exact consumers surplus and deadweight loss', *Econometrica: Journal of the Econometric Society* pp. 1445–1476.

Hirshberg, D. A. and Wager, S. (2020), 'Debiased inference of average partial effects in single-index models: Comment on wooldridge and zhu', *Journal of Business & Economic Statistics* **38**(1), 19–24.

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

Ichimura, H. and Newey, W. K. (2017), 'The influence function of semiparametric estimators', *arXiv preprint arXiv:1508.01378* .

Knittel, C. R. and Metaxoglou, K. (2012), Estimation of random coefficient demand models: Two empiricists' perspective.

Kress, R. (1989), *Linear integral equations*, Vol. 82, Springer.

Lewis, G. and Syrgkanis, V. (2018), 'Adversarial generalized method of moments', *arXiv preprint arXiv:1803.07164* .

Lu, Z., Shi, X. and Tao, J. (2019), 'Semi-nonparametric estimation of random coefficient logit model for aggregate demand', *Available at SSRN 3503560* .

Mason, L., Baxter, J., Bartlett, P. L. and Frean, M. R. (2000), Boosting algorithms as gradient descent, *in* 'Advances in neural information processing systems', pp. 512–518.

Meir, R. and Zhang, T. (2003), 'Generalization error bounds for bayesian mixture algorithms', *Journal of Machine Learning Research* **4**(Oct), 839–860.

Monardo, J. (2021), 'Measuring substitution patterns with a flexible demand model', *Available at SSRN 3921601* .

Muandet, K., Mehrjou, A., Lee, S. K. and Raj, A. (2019), 'Dual iv: A single stage instrumental variable regression', *arXiv preprint arXiv:1910.12358* .

Newey, W. K. (1994), 'The asymptotic variance of semiparametric estimators', *Econometrica: Journal of the Econometric Society* pp. 1349–1382.

Newey, W. K. (2013), 'Nonparametric instrumental variables estimation', *American Economic Review* **103**(3), 550–56.

Newey, W. K. and Powell, J. L. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Reynaert, M. and Verboven, F. (2014), 'Improving the performance of random coefficients demand models: the role of optimal instruments', *Journal of Econometrics* **179**(1), 83–98.

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

Santos, A. (2011), 'Instrumental variable methods for recovering continuous linear functionals', *Journal of Econometrics* **161**(2), 129–146.

Santos, A. (2012), 'Inference in nonparametric instrumental variables with partial identification', *Econometrica* **80**(1), 213–275.

Schapire, R. E. (1990), 'The strength of weak learnability', *Machine learning* **5**(2), 197–227.

Severini, T. A. and Tripathi, G. (2012), 'Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors', *Journal of Econometrics* **170**(2), 491–498.

Singh, R., Sahani, M. and Gretton, A. (2019), Kernel instrumental variable regression, *in* 'Advances in Neural Information Processing Systems', pp. 4595–4607.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of optimization theory and applications* **109**(3), 475–494.

Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and empirical processes: with applications to statistics*, Springer.

Van der Vaart, A. W. (1991), 'On differentiable functionals', *The Annals of Statistics* pp. 178–204.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Vershynin, R. (2018), *High-dimensional probability*, Cambridge, UK: Cambridge University Press.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural networks* **5**(2), 241–259.

Zhang, T. and Yu, B. (2005), 'Boosting with early stopping: Convergence and consistency', *The Annals of Statistics* **33**(4), 1538–1579.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429.