Automatic Debiased Machine Learning in Presence of Endogeneity†

click here for the latest version

Edvard Bakhitov‡

University of Pennsylvania

January 2, 2022

Abstract

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. This paper 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. 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. We derive a convergence rate for the penalized GMM estimator of the bias correction term. We also give conditions for root-n consistency and asymptotic normality of the debiased MLIV estimator of the functional of interest. Overall, the approach allows for a large variety of MLIV estimators as along as they satisfy mild convergence rate conditions. We apply our procedure to estimate the conditional demand derivative within the nonparametric demand for differentiated goods framework. Using both simulated and real data, we demonstrate that our debiased estimates have significantly reduced bias and close to the nominal level coverage, while the plug-in estimates perform poorly.

†I am very grateful to Amit Gandhi, Xu Cheng, and Karun Adusumilli for their guidance and support, as well as Frank Diebold, Wayne Gao, and participants of the UPenn Econometrics Workshop 2021 and Young Scholars Conference on Machine Learning in Economics and Finance 2021 for insightful comments and suggestions. I would like to thank IRI for making the data available. All estimates and analysis in this paper, based on data provided by IRI, are by the authors and not by IRI.

‡e-mail: bakhitov@sas.upenn.edu

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). 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), 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 et al. (2018a) 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 et al. (2020b), 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; Chen and Pouzo, 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 paper connects several strands of literature. First, since the focus of the paper is functionals of nonparametric quantities, our methodology relates to the literature on semi-parametric 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 our paper we focus on functionals of MLIV estimators over domains that may include low, moderate, and high dimensional objects. A more recent work by Chernozhukov et al. (2020a) generalizes and extends the insights from the classical theory by constructing Neyman orthogonal moment conditions allowing for a wide range of ML estimators[1]. We follow Chernozhukov et al. (2020a) 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 et al., 2020a). 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 et al. (2020c) 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 restrictions 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)

[1] Chernozhukov et al. (2020a) 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.

give conditions for pointwise and uniform asymptotic normality, respectively, of possibly nonlinear functionals of the sieve NPIV estimator. The results presented in this paper are complementary to the results on inference on functionals of NPIV estimators.

The paper also touches on the growing literature on flexible demand estimation in differentiated product markets. Compiani (2018) follows the nonparametric identification arguments 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. However, 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 paper is organized as follows. Section 2 briefly introduces the NPIV framework, discusses practical issues, and describes various MLIV estimators. In Section 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 4 gives conditions necessary to derive a convergence rate for the PGMM estimator. Section 5 gives conditions for root-n consistency and asymptotic normality of the debiased estimator for linear functionals. In Section 6 we introduce additional conditions necessary to extend our results to nonlinear functionals. Section 7 examines the performance of the debiased estimator in a simple Monte Carlo exercise. In Section 8 we introduce the nonparametric demand estimation framework and estimate the conditional demand derivative functional using both simulated and real data. Section 9 gives conclusions and provides possible extensions. 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 $S \subseteq \{1, \dots, n\}$ 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 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.

2.1 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,

$$\mathbb{E}[Y|Z] = \int \gamma_0(x) f(x|z) dx, \tag{1}$$

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 (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 (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 (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 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.

2.3 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). 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 A.

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

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

This is the exact problem the series NPIV estimator solves as well. Hartford et al. (2017) 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) 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 estimator 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[2]. The Dual IV (Muandet et al., 2019) uses the dual form of the NPIV estimation problem in (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.

$$\gamma_0 = \underset{\gamma \in \Gamma}{\operatorname{argmin}}\, \max_{f \in \mathcal{F}} \mathbb{E}[(Y - \gamma(X))f(Z)]. \tag{3}$$

[2] 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.

Note that the minimax problem in (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.

3 Learning functionals of MLIV estimators

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 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.

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 et al., 2020a). In the NPIV model, regularization is essential to dealing with ill-posedness rendering all NPIV/MLIV estimators regularized estimators.

Figure 1 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.

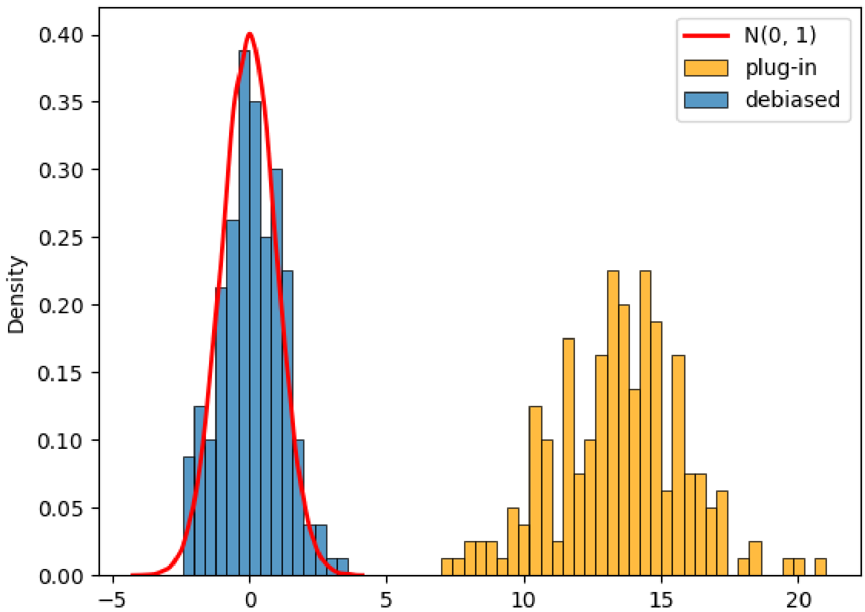

Figure 1. Distributions of plug-in and debiased estimates. The graph shows distributions of the standardized plug-in and debiased estimates of the conditional demand derivative functional from Section 8.3. The data is generated according to the logit model.

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

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

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

$$\mathbb{E}[m(W,\gamma)] = \mathbb{E}[\alpha_0(Z)\gamma(X)] \text{ for all } \gamma \text{ with } \mathbb{E}[\gamma^2(X)] < \infty. \tag{4}$$

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

$$\psi(W,\theta,\gamma,\alpha) = m(W,\gamma) - \theta + \alpha(Z)[Y - \gamma(X)], \tag{5}$$

where $\alpha(Z)[Y-\gamma(X)]$ is the influence function. Note that the moment function in (5) is Neyman orthogonal to local perturbations (δ,β) of (γ_0,α_0) such that

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

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 (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.

3.3 Estimation of the Riesz representer

Chernozhukov et al. (2020a) 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

$$\begin{aligned}
\mathbb{E}[\psi_\gamma(W,\theta_0,\delta,\alpha_0)] &= \frac{\partial}{\partial\tau}\mathbb{E}[\psi(W_i,\theta_0,\gamma_0+\tau\delta,\alpha_0)]\Big|_{\tau=0} \\
&= \frac{\partial}{\partial\tau}\mathbb{E}\left[m(W,\gamma_0+\tau\delta) - \theta_0 + \alpha_0(Z)[Y-\gamma_0(X)-\tau\delta(X)]\right]\Big|_{\tau=0} \\
&= \mathbb{E}[m(W,\delta) - \alpha_0(Z)\delta(X)] = 0, \qquad (6)
\end{aligned}$$

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 (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 et al. (2020c) 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 (6) by replacing the expectation with a sample average and $\alpha_0(Z)$ with $b(Z)'\rho$ to obtain

$$\hat{\psi}_\gamma(d_j, \rho) = \frac{1}{n}\sum_{i=1}^{n}\{m(W_i, d_j) - d_j(X_i)b(Z_i)'\rho\} = 0, \quad j = 1, \ldots, q. \tag{7}$$

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 (7). Then a solution to the PGMM problem takes the form

$$\hat{\rho}_L = \operatorname*{argmin}_{\rho \in \mathbb{R}^p} \hat{\psi}_\gamma(\rho)'\hat{\Omega}_q\hat{\psi}_\gamma(\rho) + 2\lambda_n|\rho|_1, \tag{8}$$

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 (8) in matrix form as

$$\hat{\rho}_L = \operatorname*{argmin}_{\rho \in \mathbb{R}^p}(\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda_n|\rho|_1. \tag{9}$$

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 C.

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

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

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

$$\hat{\theta} = \frac{1}{n} \sum_{\ell}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(Z_i)[Y_i - \hat{\gamma}_\ell(X_i)]\} \tag{10}$$

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

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[3]. As pointed out in Section 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).

[3]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.

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 4.

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

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

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

for some constant C.

The first part of Assumption 1 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 (9) depends on the convergence rates of $\hat{\Omega}$, $\hat{G}$, and $\hat{M}$. Assumption 1 ensures that $\hat{\Omega}$ is consistent. To obtain a convergence rate for $\hat{G}$, we impose the following condition.

Assumption 2. There are constants C_b and C_d such that with probability approaching one,

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

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 3. 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 4. 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 4 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 4 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.

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

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 5. 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 5 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 6. 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 6, $\varepsilon_n = \varepsilon_n^M = \varepsilon_n^G = \sqrt{\log(q)/n}$. 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 1. If Assumptions 2–6 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 4 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.

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 6.

We impose the following conditions.

Assumption 7. $\alpha_0(z)$ 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 8. $\int [m(w, \hat{\gamma}) - m(w, \gamma_0)]^2 F_0(dw) \xrightarrow{p} 0$ and $||\hat{\gamma} - \gamma_0|| \xrightarrow{p} 0$.

Assumption 9. $||T(\hat{\gamma} - \gamma_0)|| = O_p(\kappa_n^\gamma)$ with $\kappa_n^\gamma \to 0$.

Assumption 8 allows for estimators $\hat{\gamma}$ that are mean square consistent. Assumption 9 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 2.1). This specification is general enough and allows for various MLIV estimators.

Assumption 10. $\varepsilon_n = o(\lambda_n)$ 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 10 requires

$$\sqrt{n}\bar{s}\lambda_n\kappa_n^\gamma = O\left(n^{1/2}\left(\sqrt{\frac{\log(q)}{n}}\right)^{\frac{2\xi-1}{1+2\xi}} a_n\kappa_n^\gamma\right) \to 0. \tag{11}$$

Suppose $\kappa_n^\gamma = n^{-d_\gamma}$ with $d_\gamma > 0$. Then condition (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 2. If Assumptions 2–10 are satisfied, then for $\psi_0(w) = m(w,\gamma_0) - \theta_0 + \alpha_0(z)[y - \gamma(x)]$,

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

6 Nonlinear functionals

It is possible to extend the results from Section 5 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 et al., 2018b). 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

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

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

$$\mathbb{E}[D(W, \gamma_0, \zeta)] = \mathbb{E}[\alpha_0(Z)\zeta(X)], \text{ for all } \zeta(X) \text{ with } \mathbb{E}[\zeta^2(X)] < \infty. \tag{12}$$

In other words, Equation (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 (10) except we need a different estimator of $\alpha_0(Z)$ based on (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 et al. (2020a) for more details.

An estimator $\hat{\alpha}_\ell$ can be constructed exactly as described in Section 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

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

where $\hat{M}_{\ell j}$ is the Gateaux derivative of the moment function with respect to γ in the direction of the j^{th} 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 11. 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 1. Suppose that $||\hat{\gamma}_{\ell,\ell'} - \gamma_0|| = O_p(\kappa_n^\gamma)$ for $\ell, \ell' = 1, \ldots, L$, and Assumption 11 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 10. 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 10 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 12. 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 13. $||\hat{\gamma}-\gamma_0|| = O_p(\kappa_n^\gamma)$ 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 3. If Assumptions 2–8, 10, and 11–13 are satisfied, then for $\psi_0(w) = m(w,\gamma_0) - \theta_0 + \alpha_0(z)[y-\gamma(x)]$,

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

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

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

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 1. 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.

Table 1. 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

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 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 et al. (2020a), which indicates that our procedure is valid and performs well in practice.

8 Application to nonparametric demand estimation

8.1 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, $t=1,\dots,T$, there is a continuum of consumers choosing from a set of products $\mathcal{J}=\{0,1,\dots,J\}$ 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 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

$$\delta_{jt} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right), \quad j \in \mathcal{J}\backslash\{0\}. \tag{13}$$

We can rewrite (13) in a more convenient form to get the following estimation equation

$$x_{jt}^{(1)} = \sigma_j^{-1}\left(s_t, p_t, x_t^{(2)}\right) - \xi_{jt}. \tag{14}$$

Note that in (14) 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 (14) can be rewritten as

$$\log\left(\frac{s_{jt}}{s_{0t}}\right) = x_{jt}^{(1)} + \gamma(\omega_{jt}) + \xi_{jt}, \tag{15}$$

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 (15) in a more convenient form

$$y_{jt} = \gamma(\omega_{jt}) + \xi_{jt}. \tag{16}$$

Equation (16) 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 $z_{jt} = (\{\Delta_{jkt}^x\}_{j\neq k})$, where $\Delta_{jkt}^x = x_{jt} - x_{kt}$, 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 $z_{jt} = \left(\{\Delta_{jkt}^x, \Delta_{jkt}^c\}_{j\neq k}\right)$, where $\Delta_{jkt}^c = c_{jt} - c_{kt}$.

8.2 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 (15),

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

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 (15) 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 (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.

8.3 Simulated data experiments

8.3.1 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 (15) 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 $j = 1, \dots, J$ and $t = 1, \dots, T$,

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

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[4] 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 2. 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 2 presents the results. We can clearly see the bias-variance trade-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

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

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.

8.3.2 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

$$\theta_{jj} = \mathbb{E}\left[\beta_p + \frac{\pi}{s_{j|gt}}\frac{\partial s_{j|gt}}{\partial p_{jt}}\right]. \tag{17}$$

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,

$$\frac{\partial s_{j|gt}}{\partial p_{jt}} = \frac{\partial s_{j|gt}}{\partial \delta_{jt}}\frac{\partial \delta_{jt}}{\partial p_{jt}} = \frac{\beta_p}{1-\pi}s_{j|gt}(1 - s_{j|gt}). \tag{18}$$

Thus, combining (17) and (18) gives

$$\theta_{jj} = \mathbb{E}\left[\beta_p\left(1 + \frac{\pi}{1-\pi}(1 - s_{j|gt})\right)\right] = \underbrace{\beta_p}_{\text{direct effect}} + \underbrace{\beta_p\frac{\pi}{1-\pi}\mathbb{E}[1 - s_{j|gt}]}_{\text{indirect effect}}. \tag{19}$$

Note that in equation (19) 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, $x_{3,jt}^{(2)}$. 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 3. 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 3. 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.

8.4 Real data example

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 E 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 4 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 (19). 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 basis functions as in GNT (see Appendix F 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 5 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 (19) 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

Table 4. 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.

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 is more sensitive.

We can see a clear debiasing effect in the last two columns of Table 5. 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 Section 8.3.

Table 5. 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
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
Pepsi	-5.340	0.193	-5.046	0.045	-6.172	0.302
Sprite	-5.539	0.060	-4.015	0.020	-6.025	0.552

9 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.

Finally, we have applied our debiasing procedure 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.

A 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.

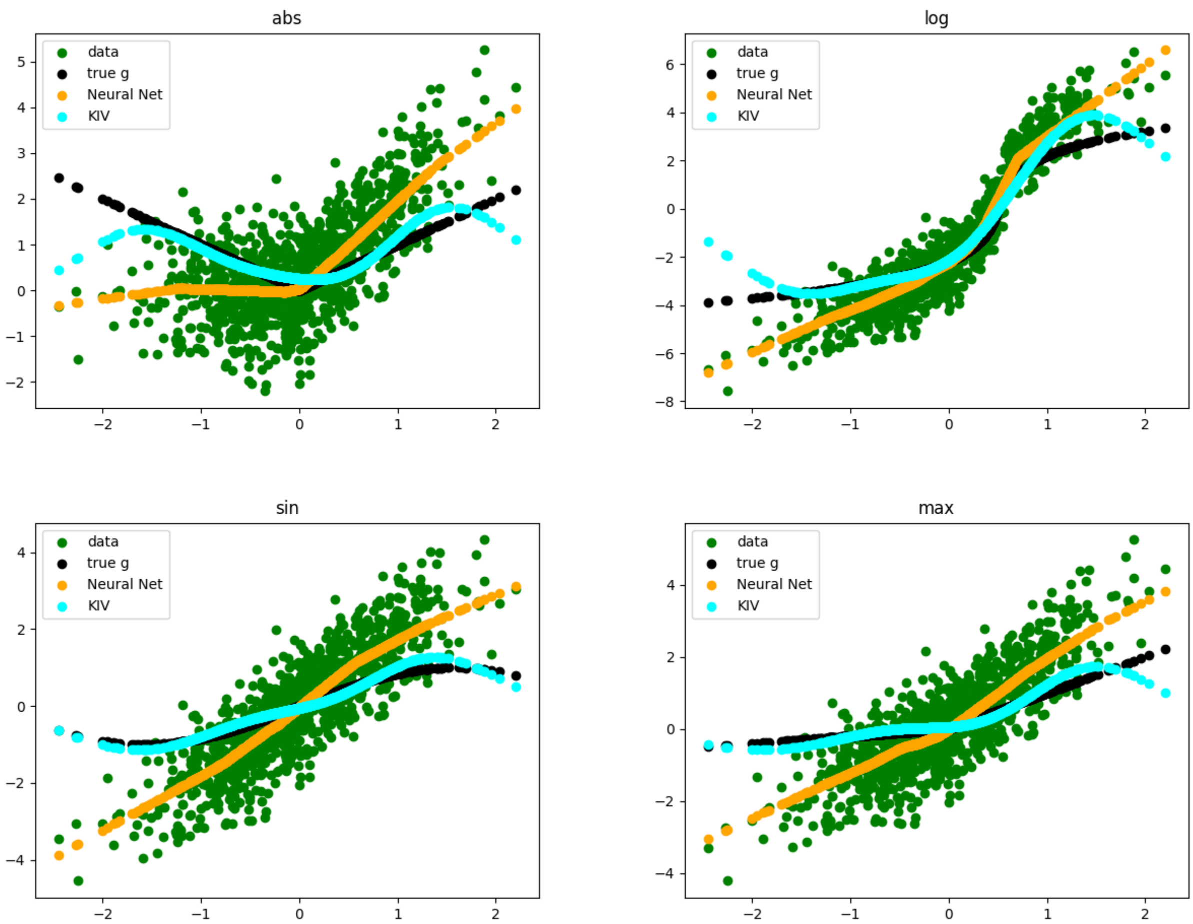

Figure 2. Standard ML vs MLIV estimators

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,

abs: $\gamma(X) = |X|$, **log:** $\gamma(X) = \log(|16X - 8| + 1)\text{sign}(X - 0.5)$

sin: $\gamma(X) = \sin(X)$, **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 2 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 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

$$\min_{\rho \in \mathbb{R}^p} (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho). \tag{B.1}$$

Given the form of the debiased moment function (5) and the linear approximation for the RR, the orthogonal moment condition (7) will always be linear in ρ, meaning that the GMM criterion in (B.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

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

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

$$\frac{\partial \hat{\psi}_\gamma(\hat{\rho})}{\partial \rho'}\hat{\Omega}_q \left\{ \frac{1}{n}\sum_{i=1}^{n} m(W_i, d) - \frac{1}{n}\sum_{i=1}^{n} d(X_i)b(Z_i)'\hat{\rho} \right\} = 0. \tag{B.2}$$

We can rewrite (B.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 $\hat{\rho}$,

$$\hat{\rho} = (\hat{G}'\hat{\Omega}_q\hat{G})^{-1}\hat{G}'\hat{\Omega}_q\hat{M}. \tag{B.3}$$

Note that the form of the GMM solution in (B.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.

C Computing Auto-DML using Penalized GMM

Recall, in matrix form the PGMM problem is given by

$$\min_{\rho \in \mathbb{R}^p} (\hat{M} - \hat{G}\rho)'\hat{\Omega}_q(\hat{M} - \hat{G}\rho) + 2\lambda_n|\rho|_1. \tag{C.1}$$

Note that the objective in (C.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 j^{th} element of a generic vector v by v_j and let e_j be a $p \times 1$ unit vector with 1 in the j^{th} coordinate and zeros elsewhere.

Algorithm 1 Coordinate-wise descent algorithm for PGMM

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

2: Calculate loadings that do not depend on ρ_j:

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

3: Update coordinate ρ_j:

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

4: **end for**

The justification for Algorithm 1 is similar to the one of CNS. It follows from the fact that the GMM objective (C.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 (C.1) into the adaptive PGMM (A-PGMM) problem

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

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 1 is that in step 3 we replace λ_n with $\hat{w}_j\lambda_n$.

C.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.

C.1.1 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[5].

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[6].

In Table 6, 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 adding adaptive weights on the penalty term improves the performance twice rendering the lowest MSE across the algorithms, which validates the procedure.

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

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

Table 6. **HD Linear regression results.**

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

C.1.2 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 7 demonstrates MSEs of the considered implementations based on 200 simulations.

Table 7. **HD Linear IV regression results.**

	MSE
Double Lasso	0.1864
PGMM	0.3020
A-PGMM	0.0726

D 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.

D.1 Properties of the PGMM estimator

Lemma D.1. If Assumption 2 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 et al. (2020b). 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,

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

where the last inequality follows from $q \geq p$. Note that $\mathbb{E}[T_{ijk}] = 0$ and by Assumption 2,

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

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 D.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

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

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 et al. (2020b). 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.

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

Thus, for $j = 1, \dots, p$ 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 j^{th} 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 4 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 $\bar{\rho} = (\bar{\rho}_1, \dots, \bar{\rho}_p)'$, 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

$$\rho_\star = \underset{v \in \mathbb{R}^p}{\text{argmin}} (\rho_L - v)'G'\Omega_q G(\rho_L - v) + 2\varepsilon_n \sum_{j \in S_{\bar{\rho}}^c} |v_j|. \tag{D.1}$$

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

Lemma D.3. $||G'\Omega_q G(\rho_\star - \rho_L)||_\infty \leq \varepsilon_n$.

Proof. Follows directly from the proof of Lemma D.2. ■

Lemma D.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 D.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 (D.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 D.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 D.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 4,

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

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 D.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 D.7. If Assumptions 1–3 and 6 are satisfied, then

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

Proof. By the triangle inequality,

$$\begin{aligned} ||\hat{G}'\hat{\Omega}_q(\hat{M} - \hat{G}\rho_\star)||_\infty \leq\; & ||\hat{G}'\hat{\Omega}_q\hat{M} - G'\Omega_q M||_\infty & \text{(D.2)} \\ & + ||G'\Omega_q M - G'\Omega_q G\rho_\star||_\infty & \text{(D.3)} \\ & + ||(G'\Omega_q G - \hat{G}'\hat{\Omega}_q\hat{G})\rho_\star||_\infty. & \text{(D.4)} \end{aligned}$$

Consider the first element (D.2). Note that by the triangle inequality,

$$\begin{aligned} ||\hat{G}'\hat{\Omega}_q\hat{M} - G'\Omega_q M||_\infty \leq\; & ||(\hat{G} - G)'(\hat{\Omega}_q - \Omega_q)(\hat{M} - M)||_\infty & \text{(D.5)} \\ & + ||(\hat{G} - G)'\Omega_q(\hat{M} - M)||_\infty & \text{(D.6)} \\ & + ||G'(\hat{\Omega}_q - \Omega_q)(\hat{M} - M)||_\infty & \text{(D.7)} \\ & + ||G'\Omega_q(\hat{M} - M)||_\infty & \text{(D.8)} \\ & + ||(\hat{G} - G)'\Omega_q M||_\infty & \text{(D.9)} \\ & + ||(\hat{G} - G)'(\hat{\Omega}_q - \Omega_q)M||_\infty & \text{(D.10)} \\ & + ||G'(\hat{\Omega}_q - \Omega_q)M||_\infty. & \text{(D.11)} \end{aligned}$$

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

$$||BFA||_\infty \leq q||B||_\infty||F||_{\ell_\infty}||A||_\infty. \tag{D.12}$$

We can use (D.12) to put an upper bound on (D.5),

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

Moreover, notice that Assumptions 2 and 6 imply that $||G||_\infty = O(1)$ and $||M||_\infty = O(1)$. Using this fact and (D.12), we can bound the remaining terms (D.6)–(D.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 (D.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 D.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 D.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 (D.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 (D.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 (D.2)–(D.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 5 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 1.

Lemma D.8. If Assumptions 2 and 1 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

$$|x'BFAx| \leq q|x|_1^2||B||_\infty||F||_{\ell_\infty}||A||_\infty. \tag{D.13}$$

Using (D.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

$$\begin{aligned}|\delta'\hat{G}'\hat{\Omega}_q\hat{G}\delta| \geq |\delta'G'\Omega_qG\delta| \\ &- |\delta|_1^2||\hat{G}-G||_\infty^2(||\hat{\Omega}-\Omega||_{\ell_\infty}+||\Omega||_\infty) \\ &- |\delta|_1^2||G||_\infty^2||\hat{\Omega}-\Omega||_{\ell_\infty} \\ &- 2|\delta|_1^2||\hat{G}-G||_\infty||G||_\infty(||\hat{\Omega}-\Omega||_{\ell_\infty}+||\Omega||_\infty)\end{aligned} \quad \text{(D.14)}$$

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

$$|\delta|_1 \leq 4\sqrt{s_\star}||\delta_{S_{\rho_\star}}|| \Rightarrow \frac{|\delta|_1^2}{||\delta_{S_{\rho_\star}}||^2} \leq 16s_\star. \quad \text{(D.15)}$$

Divide (D.14) by $||\delta_{S_{\rho_\star}}||^2$ and use (D.15),

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

Since $\frac{|\delta'G'\Omega_qG\delta|}{||\delta_{S_{\rho_\star}}||^2} \geq \phi^2(s_\star)$ 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

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

Using Assumptions 2 and 1 and Lemma D.1, we can put an upper bound on a_n as follows

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

Gathering the terms gives

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

which completes the proof. ■

Proof of Theorem 1

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

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

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

$$||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\hat{\rho}_L)||^2+2\lambda_n|\hat{\rho}_L|_1 \le ||\hat{\Omega}_q^{1/2}(\hat{M}-\hat{G}\rho_\star)||^2+2\lambda_n|\rho_\star|_1. \tag{D.16}$$

First, observe that

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

Plug the expression above in (D.16) to get

$$\begin{aligned}
||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L-\rho_\star)||^2+2\lambda_n|\hat{\rho}_L|_1 &\le 2(\hat{G}'\hat{\Omega}_q\hat{M}-\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star)'(\hat{\rho}_L-\rho_\star)+2\lambda_n|\rho_\star|_1\\
&\le 2||\hat{G}'\hat{\Omega}_q\hat{M}-\hat{G}'\hat{\Omega}_q\hat{G}\rho_\star||_\infty|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1\\
&= 2||\hat{G}'\hat{\Omega}_q(\hat{M}-\hat{G}\rho_\star)||_\infty|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1\\
&= 2o_p(\lambda_n)|\hat{\rho}_L-\rho_\star|_1+2\lambda_n|\rho_\star|_1,
\end{aligned} \tag{D.17}$$

where the second inequality is Hölder and the last equality comes from Lemma D.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

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + \lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq 3\lambda_n|\hat{\rho}_{L,S_{\rho_\star}} - \rho_{\star,S_{\rho_\star}}|_1. \tag{D.18}$$

The inequality in (D.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 (D.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

$$||\hat{\Omega}_q^{1/2}\hat{G}'(\hat{\rho}_L - \rho_\star)||^2 + 2\lambda_n|\hat{\rho}_{L,S^c_{\rho_\star}}|_1 \leq \frac{9\lambda_n^2 s_\star}{\hat{\phi}^2(s_\star)}. \tag{D.19}$$

To get the ℓ_1-error bound, ignore the first term on the LHS of (D.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 (D.19) implies

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

Next, by Lemma D.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 D.5 gives

$$|\hat{\rho}_L - \rho_\star|_1 = O_p(\bar{s}\lambda_n). \tag{D.20}$$

Finally, let $\alpha_\star = b(Z)'\rho_\star$, then by the triangle inequality and Lemma D.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 (D.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 $\bar{s}$.

■

D.2 Asymptotic properties

Lemma D.9. If Assumptions 1–3 and 6 are satisfied and $\varepsilon_n = o(\lambda_n)$, then

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

Proof. Recall Equation (D.17) from the proof of Theorem 1 which implies

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

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

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

which implies that with probability approaching one,

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

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

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

■

Proof of Theorem 2

We prove the first conclusion by verifying the conditions of Lemma 15 of Chernozhukov et al. (2020a). Let $g(w,\gamma,\alpha,\theta)$ and $\phi(w,\gamma,\alpha,\theta)$ in Chernozhukov et al. (2020a) 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 7. Moreover, note that by Assumptions 7 and 8, Theorem 1, and the law of iterated expectations,

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

Also, it follows from Assumption 8 that

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

Thus, all the conditions of Assumption 1 of Chernozhukov et al. (2020a) are 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 7 and $\sup_z |\hat{\alpha}_\ell(z)| = O_p(1)$ by Lemma D.9,

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

where the conclusion follows from Assumption 8. Furthermore, by Cauchy-Schwarz inequality and Assumption 10,

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

which renders Assumption 2(iii) of Chernozhukov et al. (2020a) satisfied.

Also, by construction,

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

Combined with $m(w,\gamma)$ being affine in γ verifies Assumption 3 of Chernozhukov et al. (2020a) 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 et al. (2020b). 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,

$$\hat{V} - V = \frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i - \psi_i)^2 + \frac{2}{n}\sum_{i=1}^n(\hat{\psi}_i - \psi_i)\psi_i \leq \frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i - \psi_i)^2 + 2\sqrt{\frac{1}{n}\sum_{i=1}^n(\hat{\psi}_i - \psi_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \psi_i^2}. \tag{D.21}$$

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 2,

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

By Lemma D.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 7, 8, 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 7 and 8. Thus, the conclusion follows from (D.21) and the central limit theorem.

■

Proof of Lemma 1

The proof is similar to the proof of Lemma 10 of Chernozhukov et al. (2020b). 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 11. When $\Gamma_{\ell,\ell'}$ occurs,

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

by Assumption 11. 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 $\varepsilon_n = \sqrt{\log(q)/n}$,

$$\begin{aligned} \mathbb{P}(\mathcal{A}) &= \mathbb{P}(\mathcal{A}|\Gamma_{\ell,\ell'})\mathbb{P}(\Gamma_{\ell,\ell'}) + \mathbb{P}(\mathcal{A}|\Gamma_{\ell,\ell'}^c)[1 - \mathbb{P}(\Gamma_{\ell,\ell'})] \\ &\leq \mathbb{P}\left(\max_{1 \leq j \leq q} |U_{ij}(\tilde{\gamma}_{\ell,\ell'})| \geq \bar{C}\varepsilon_n | \Gamma_{\ell,\ell'}\right) + [1 - \mathbb{P}(\Gamma_{\ell,\ell'})]. \end{aligned} \tag{D.22}$$

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 D.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

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

for any $\bar{C}>K\sqrt{L/c}$. Let $U_{\ell'}(\gamma)=(U_{\ell'1},\ldots,U_{\ell'q})'$. Then it follows from (D.22) that for large $\bar{C}$, $\mathbb{P}(|U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})|\ge\bar{C}\varepsilon_n)\to 0$, meaning that $||U_{\ell'}(\tilde{\gamma}_{\ell,\ell'})||_\infty=O_p(\varepsilon_n)$.

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

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

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

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

Also, by Assumption 11(ii) and $\mathbb{P}(\Gamma_{\ell,\ell'})\to 1$ for each ℓ and ℓ',

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

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

Proof of Theorem 3

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

Next, by Assumptions 12 and 13,

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

Moreover, as in the proof of Theorem 2,

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

Thus, all the conditions of Assumption 3 of Chernozhukov et al. (2020a) are satisfied, which combined with the results above gives us the first conclusion. The second conclusion follows exactly as in the proof of Theorem 2. ■

E Data cleaning and aggregation details

E.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.

E.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”, ”WILD 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

F GNT basis functions

Here we present an idea behind the approximation strategy in GNT. We have a function $\gamma\left(\omega_{jt}\right)$ 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 symmetric 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 n^{th} 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 n^{th} order approximation for each t is

$$\bigcup_{i=2}^{n} B_i^{jt}$$

The choice of g can be determine by any functional form that allows for a flexible approximation 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 $q = 594$.

References

Ackerberg, Daniel, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao (2014). "Asymptotic efficiency of semiparametric two-step GMM". In: *Review of Economic Studies* 81(3), pp. 919–943.

Ai, Chunrong and Xiaohong Chen (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions". In: *Econometrica* 71(6), pp. 1795–1843.

Ai, Chunrong and Xiaohong Chen (2007). "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables". In: *Journal of Econometrics* 141(1), pp. 5–43.

Bakhitov, Edvard, Amit Gandhi, and Jing Tao (2020). *Feature Selection in Differentiated Product Demand Models*. Tech. rep. Working paper.

Bakhitov, Edvard and Amandeep Singh (2021). "Causal Gradient Boosting: Boosted Instrumental Variable Regression". In: *arXiv preprint arXiv:2101.06078*.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012). "Sparse models and methods for optimal instruments with an application to eminent domain". In: *Econometrica* 80(6), pp. 2369–2429.

Bennett, Andrew, Nathan Kallus, and Tobias Schnabel (2019). "Deep generalized method of moments for instrumental variable analysis". In: *Advances in Neural Information Processing Systems*, pp. 3559–3569.

Berry, S., J. Levinsohn, and A. Pakes (1995). "Automobile prices in market equilibrium". In: *Econometrica: Journal of the Econometric Society*, pp. 841–890.

Berry, Steven T and Philip A Haile (2014). "Identification in differentiated products markets using market level data". In: *Econometrica* 82(5), pp. 1749–1797.

Berry, Steven, Amit Gandhi, and Philip Haile (2013). "Connected substitutes and invertibility of demand". In: *Econometrica* 81(5), pp. 2087–2111.

Berry, Steven and Philip Haile (2016). "Identification in differentiated products markets". In: *Annual review of Economics* 8, pp. 27–52.

Bickel, Peter J, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore.

Bickel, Peter J, Ya'acov Ritov, Alexandre B Tsybakov, et al. (2009). "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of statistics* 37(4), pp. 1705–1732.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen (2007). "Semi-nonparametric IV estimation of shape-invariant Engel curves". In: *Econometrica* 75(6), pp. 1613–1669.

Bradic, Jelena, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu (2021). "Minimax semiparametric learning with approximate sparsity". In: *arXiv preprint arXiv:1912.12213*.

Bronnenberg, Bart J, Michael W Kruger, and Carl F Mela (2008). "Database paper—The IRI marketing data set". In: *Marketing science* 27(4), pp. 745–748.

Caner, Mehmet and Anders Bredahl Kock (2018). "High Dimensional Linear GMM". In: *arXiv preprint arXiv:1811.08779*.

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault (2007). "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization". In: *Handbook of econometrics* 6, pp. 5633–5751.

Chen, Jiafeng, Xiaohong Chen, and Elie Tamer (2021). "Efficient Estimation in NPIV Models: A Comparison of Various Neural Networks-Based Estimators". In: *arXiv preprint arXiv:2110.06763*.

Chen, Xiaohong and Timothy M Christensen (2018). "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression". In: *Quantitative Economics* 9(1), pp. 39–84.

Chen, Xiaohong and Demian Pouzo (2012). "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals". In: *Econometrica* 80(1), pp. 277–321. ISSN: 1468-0262. DOI: 10.3982/ECTA7888. URL: http://dx.doi.org/10.3982/ECTA7888.

Chen, Xiaohong and Demian Pouzo (2015). "Sieve Wald and QLR inferences on semi/nonparametric conditional moment models". In: *Econometrica* 83(3), pp. 1013–1079.

Chernozhukov, V, Whitney K Newey, James Robins, and Rahul Singh (2019). "Double/de-biased machine learning of global and local parameters using regularized Riesz representers". In: *stat* 1050, p. 9.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins (2020a). "Locally robust semiparametric estimation". In: *arXiv preprint arXiv:1608.00033*.

Chernozhukov, Victor, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis (2021). "Automatic Debiased Machine Learning via Neural Nets for Generalized Linear Regression". In: *arXiv preprint arXiv:2104.14737*.

Chernozhukov, Victor, Whitney K Newey, and James Robins (2020b). *Automatic debiased machine learning of causal and structural effects*. Tech. rep.

Chernozhukov, Victor, Whitney K Newey, and James Robins (2018a). *Double/de-biased machine learning using regularized Riesz representers*. Tech. rep. cemmap working paper.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (2018b). "Learning L2 continuous regression functionals via regularized Riesz representers". In: *arXiv preprint arXiv:1809.05224*.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis (2020c). "Adversarial Estimation of Riesz Representers". In: *arXiv preprint arXiv:2101.00009*.

Compiani, Giovanni (2018). "Nonparametric demand estimation in differentiated products markets". In: *Available at SSRN 3134152*.

Conlon, Christopher and Jeff Gortmaker (2020). "Best practices for differentiated products demand estimation with pyblp". In: *The RAND Journal of Economics* 51(4), pp. 1108–1161.

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault (2011). "Nonparametric instrumental regression". In: *Econometrica* 79(5), pp. 1541–1565.

Dikkala, Nishanth, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis (2020). "Minimax estimation of conditional moment models". In: *arXiv preprint arXiv:2006.07201*.

Fosgerau, Mogens, Julien Monardo, and André De Palma (2020). "The inverse product differentiation logit model". In: *Available at SSRN 3141041*.

Friedman, Jerome, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. (2007). "Pathwise coordinate optimization". In: *The annals of applied statistics* 1(2), pp. 302–332.

Gandhi, Amit and Jean-François Houde (2019). *Measuring substitution patterns in differentiated products industries*. Tech. rep. National Bureau of Economic Research.

Gandhi, Amit, Aviv Nevo, and Jing Tao (2020). *Flexible Estimation of Differentiated Product Demand Models Using Aggregate Data*. Tech. rep. Working paper.

Gautier, Eric and Christiern Rose (2021). "High-dimensional instrumental variables regression and confidence sets". In: *arXiv preprint arXiv:1105.2454*.

Gold, David, Johannes Lederer, and Jing Tao (2020). "Inference for high-dimensional instrumental variables regression". In: *Journal of Econometrics* 217(1), pp. 79–111.

Hall, Peter and Joel L Horowitz (2005). "Nonparametric methods for inference in the presence of instrumental variables". In: *The Annals of Statistics* 33(6), pp. 2904–2929.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy (2017). "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, Jerry A and Whitney K Newey (1995). "Nonparametric estimation of exact consumers surplus and deadweight loss". In: *Econometrica: Journal of the Econometric Society*, pp. 1445–1476.

Hirshberg, David A and Stefan Wager (2020). "Debiased Inference of Average Partial Effects in Single-Index Models: Comment on Wooldridge and Zhu". In: *Journal of Business & Economic Statistics* 38(1), pp. 19–24.

Horowitz, Joel L (2011). "Applied nonparametric instrumental variables estimation". In: *Econometrica* 79(2), pp. 347–394.

Ichimura, Hidehiko and Whitney K Newey (2017). "The influence function of semiparametric estimators". In: *arXiv preprint arXiv:1508.01378*.

Kress, Rainer (1989). *Linear integral equations*. Vol. 82. Springer.

Lewis, Greg and Vasilis Syrgkanis (2018). "Adversarial generalized method of moments". In: *arXiv preprint arXiv:1803.07164*.

Lu, Zhentong, Xiaoxia Shi, and Jing Tao (2019). "Semi-Nonparametric Estimation of Random Coefficient Logit Model for Aggregate Demand". In: *Available at SSRN 3503560*.

Monardo, Julien (2021). "Measuring substitution patterns with a flexible demand model". In: *Available at SSRN 3921601*.

Muandet, Krikamol, Arash Mehrjou, Si Kai Lee, and Anant Raj (2019). "Dual IV: A Single Stage Instrumental Variable Regression". In: *arXiv preprint arXiv:1910.12358*.

Newey, Whitney K (2013). "Nonparametric instrumental variables estimation". In: *American Economic Review* 103(3), pp. 550–56.

Newey, Whitney K (1994). "The asymptotic variance of semiparametric estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Newey, Whitney K and James L Powell (2003). "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71(5), pp. 1565–1578.

Reynaert, Mathias and Frank Verboven (2014). "Improving the performance of random coefficients demand models: the role of optimal instruments". In: *Journal of Econometrics* 179(1), pp. 83–98.

Robins, James M and Andrea Rotnitzky (1995). "Semiparametric efficiency in multivariate regression models with missing data". In: *Journal of the American Statistical Association* 90(429), pp. 122–129.

Santos, Andres (2012). "Inference in nonparametric instrumental variables with partial identification". In: *Econometrica* 80(1), pp. 213–275.

Santos, Andres (2011). “Instrumental variable methods for recovering continuous linear functionals”. In: *Journal of Econometrics* 161(2), pp. 129–146.

Severini, Thomas A and Gautam Tripathi (2012). “Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors”. In: *Journal of Econometrics* 170(2), pp. 491–498.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2019). “Kernel instrumental variable regression”. In: *Advances in Neural Information Processing Systems*, pp. 4595–4607.

Tseng, Paul (2001). “Convergence of a block coordinate descent method for nondifferentiable minimization”. In: *Journal of optimization theory and applications* 109(3), pp. 475–494.

Van der Vaart, Aad W (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Van der Vaart, Aad W (1991). “On differentiable functionals”. In: *The Annals of Statistics*, pp. 178–204.

Vershynin, Roman (2018). *High-dimensional probability*. Cambridge, UK: Cambridge University Press.

Zou, Hui (2006). “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101(476), pp. 1418–1429.