AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n° 92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 1 Capitole

Présentée et soutenue par

Hippolyte BOUCHER

Le 19 avril 2023

Essais sur les Tests de Spécification et la Sélection de Modèles à Variables Instrumentales

Ecole doctorale : **TSE - Toulouse Sciences Economiques**

Spécialité : **Sciences Economiques - Toulouse**

Unité de recherche :
TSE-R - Toulouse School of Economics - Recherche

Thèse dirigée par
Pascal LAVERGNE

Jury

M. Frank WINDMEIJER, Rapporteur
Mme Bertille ANTOINE, Rapporteure
M. Eric GAUTIER, Examinateur
M. Pascal LAVERGNE, Directeur de thèse

PhD Thesis in Economics

Essays on Specification Testing and Model Selection in Instrumental Variable Models

Hippolyte Boucher

April 19, 2023

Acknowledgements

I am beholden to my advisor Pascal Lavergne and I wish to convey my profound gratitude and respect to him. He gave me great ideas and supplemented and improved mine, he shared his numerous methods and skills, he taught me rigor and how to present my work, most importantly he was always very patient, calming and supportive and helped me navigate my PhD during these four years. Next, I must express my deep appreciation to Eric Gautier and Frank Windmeijer for taking their time to share their knowledge and discuss with me. Their comments and suggestions were extremely helpful and had a tremendous influence on my own research. In addition, Frank invited me to the Oxford University Economics department, which allowed me to get invaluable feedback, especially on my job market paper, I am extremely grateful. Along with Eric and Frank, I also want to thank Bertille Antoine for being my jury members.

I feel very privileged to be friends and co-authors with Max Lesellier and Gökçe Gökkoca. They are brilliant researchers and fantastic people with whom I shared this journey. I also want to thank some of my fellow PhD students and friends most notably Young Kim, Anaïs Fabre, Tim Ederer, Tuuli Vanhapelto, Lisa Botbol, Léa Bignon, Peter Neis, Jose Alfonso Munoz Alvarado, and Mortiz Loewenfeld. We shared many things together, we had amazing discussions, they helped me do better research, broadened my horizons and made me grow as a person. I had a great time thanks to them and wish them the very best. I am also thankful to my colleagues, fellow economists and friends from Toulouse School of Economics and beyond for their help with my research, their support when teaching, with their life advice, or simply by making my life as a PhD student easier, among others Nour Meddahi, Jean-Pierre Florens, Sylvain Chabé-Ferret, François Poinas, Koen Jochmans, Sophocles Mavroeidis, Jad Beyhum, Christophe Gaillac, Alipio Ferreira, Luise Eisfeld, Claudia Noack, Louise Strachan and Audin Roger.

Lastly, I give many thanks to my family and non-economist friends for their continuous support, especially my wife Thao. We married at the beginning of this journey, she was always supportive and showered me with love, she is a most splendid life-partner and I cannot wait to walk through life's next steps with her. This thesis is dedicated to her.

Overview

This thesis contains four chapters on specification tests and model selection of instrumental variable models. Instrumental variables (IVs) have become a major tool in the social sciences and in the evaluation of public policies as they allow researchers to estimate causal effects of endogenous variables on outcomes without bias (for instance due to omitted variables or measurement error) by projecting said endogenous variables on the IVs. In order for this strategy to work IVs must be correlated with the endogenous variables and must be exogenous, ie the IVs must affect the outcome only through the endogenous variables. Thus, in this thesis, I design tests and methods to choose IV models in order to estimate the true causal effect of one variable on another. In each chapter I focus on a specific class of commonly used IV model and then develop methods either to test whether the model is well-specified or not, or to select the "best" model, or both. This thesis is a nice contribution because the statistics and econometrics literature has mainly focused on choosing the best model for prediction or finding the best causal model in simple cases (linear model, binary classifier). At the same time, a deliberate decision was to make these new methods easily implementable thus usable by applied researchers, even tough they are very technical. Below is a description of each chapter for non-specialists, the table of contents, and the chapters themselves.

The first chapter is called "A Pivotal Nonparametric Test for Identification-Robust Inference in Linear IV Models" and largely builds upon the ideas developed in Antoine and Lavergne (2019). In it, I consider the linear IV model with homogenous effect and independent data (which is the most common type of IV model) and develop a testing method for the values of the causal effect. As a direct consequence, by testing and rejecting different values, confidence intervals of the causal effect can be built. This test has three notable characteristics. First, it is robust to identification problems, in other words even if the true causal effect cannot be consistently estimated, specific values of this causal effect can be properly rejected with the test leading to valid confidence intervals. Second, the test is nonparametric and utilizes all the information contained in the relationship between the endogenous variable and the IVs. In turn, this makes the confidence interval for the causal effect built using the test small. Third, the test statistic is pivotal as in its asymptotic distribution is chi-square distributed under the null which makes it very simple to apply in practice. These characteristics are demonstrated in a simulation exercise and in an application. In the latter, I use the data from Angrist and Krueger (1991) and show, using numerous IVs, that the true causal effect of education on wages is significant and positive. Although this procedure can be extended to the time series

case, it requires either to impose very strong restriction on the autocorrelation structure or to change the test statistic (and remove its pivotality which is its main appeal).

The second chapter is called "Testing and Relaxing Distributional Assumptions on Random Coefficients in Demand Models" and is co-authored with Gökçe Gökkoca and Max Lesellier (it is his job market paper). This large paper is the product of our understanding of the econometrics literature on tests and misspecification, and of our understanding of the empirical industrial organization (IO) literature and some of its failings. In it, we consider the differentiated products demand model of Berry (1994) and Berry, Levinsohn and Pakes (1995) also known as the BLP model. The BLP model is the staple model in the empirical industrial IO literature as it only requires macro-level data, with variables such as price which can be endogenous (thus the model has IVs), while allowing for individual preference heterogeneity. Preference heterogeneity is key and determines product substitution patterns which are then used to perform counterfactual exercises (for instance estimating the effect of a new carbon tax on demand for cars). The most common way to introduce preference heterogeneity consists in introducing normally distributed random coefficients. But although this method is simple, it may not capture the true preferences of individuals leading to the wrong substitution patterns, thus the wrong counterfactuals. In this chapter, we show that choosing the wrong distribution for the random coefficients leads to a misspecification error which is correlated with the IVs. Then we develop a class of specification tests of the distribution random coefficients based on this correlation. In addition, we build new IVs called most powerful instruments, which take into this potential misspecification to maximize the tests' power and to better estimate the BLP model. Furthermore, we propose an algorithm to determine which product attributes should display preference heterogeneity. The empirical performances of the tests, of the most powerful instruments, and of the algorithm are assessed with simulations, and we show, using data from the German car market, that a flexible distribution of preferences leads to very different demand functions and counterfactual quantities of interest. Note that the tests and the new IVs created in this paper generalize to other models and frameworks, and thus will soon become their own separate papers.

The third chapter of this thesis and also my job market paper is called "Selecting Strong and Exogenous Instruments via Structural Error Criteria". This paper is the result of my insights in the econometrics literature on weak IVs and invalid IVs, and of my exposure to the statistics and machine learning literature on model selection. I consider the linear IV model with homogenous effect and independent and identically distributed data, but I also

allow some of the IVs to have a direct linear effect on the dependent variable and / or to have a low correlation with the endogenous variable. From there, only a subset of IVs can be used to estimate consistently the true causal effect of the endogenous variable. A typical example is when estimating the effect of price on demand and using weather variables as IVs. Weather variables are cost-shifters, however, they may be weak or may directly affect demand. Consequently, I coin three relevant losses and criteria based on out-of-sample validation to select the subsets of IVs which correctly estimate the causal effect. Then I provide theoretical and simulation evidence that the correct IV subsets are selected and that the true causal effect is estimated. In the application, I estimate that the causal effect of pre-trial detention on the probability of being found guilty is 25% after selecting judge dummy IVs (in comparison to 18% without selection). Note that although these selection procedures and their good properties generalize to the heteroskedastic case and to various types of IV estimators, their empirical performances can falter when the model is not linear or when the causal effect is heterogeneous. This makes their interpretation and practical use difficult. Thus, the next step in this project is to derive the theoretical properties of the selection criteria designed in this paper for more general models.

The fourth and last chapter is the vignette of the package "SpeTestNP" developed on R in collaboration with Pascal Lavergne. This project started early during my PhD as a simple package and was finally published on CRAN (the official repertory for all R libraries) in 2022. This package performs nonparametric tests of parametric specifications. In simpler terms, it allows to test whether a simple (parametric) model is capable of capturing the full effect of one variable on another compared to a more complex (non-linear) model. A very large class of parametric models can be tested (including IV models if the functions are well-parametrized). Five heteroskedasticity-robust tests are available: Bierens (1982), Zheng (1996), Escanciano (2006), Lavergne and Patilea (2008), and Lavergne and Patilea (2012). Specific bandwidth and kernel methods can be chosen, along with many other options, most notably parallel computing to quickly compute p-values based on the bootstrap. The package capabilities are illustrated by testing parametric specifications of the effect of years of education and age on earnings.

Contents

Chapter 1: A Pivotal Nonparametric Test for Identification-Robust Inference in Linear IV Models

Abstract

In linear models with endogenous regressors it is well-known that weak instruments (IVs) bias the 2 Stage Least Squares (2SLS) and other k-class IV estimators and make standard Gaussian confidence intervals invalid. Inference can still be performed by inverting tests, however there are no known method to account for a non-linear first stage except Antoine and Lavergne (2022). Their method requires simulations of the distribution of the test statistic under the null which makes it difficult to apply when sample size is moderate to large. For the above reasons I build a pivotal test statistic based on a score of integrated conditional moments which allows to easily infer on the model's structural parameters regardless of instruments' strength and the shape of the first stage conditional mean. For heteroskedastic or independent and identically distribution data with normal or non-normal errors I prove that the test is valid regardless of the degree of identification of the structural parameter of interest, and also prove that the test is consistent as long if the parameter of interest is at least semi-strongly identified. I compare the performances of the test against competing ones and revisit the effect of education on wage using Angrist and Krueger (1991) data and prove that it is strictly positive.

Keywords: Weak Instruments, Hypothesis Testing, Semiparametric Model
JEL Codes: C12, C13, C14

1 Introduction

Consider the linear model with endogenous variables popular in reduced form econometrics

$$y_i = x_i'\beta + z_{1i}'\gamma + u_i \qquad \mathbb{E}(u_i|z_{1i}, z_{2i}) = 0 \quad i = 1, \ldots, n \tag{1.1}$$

where x are endogenous variables, z_1 are exogenous control variables, and z_2 are exogenous instrumental variables. One would like to infer on the structural parameter β. Starting from a controversial application by Angrist and Krueger (1991) and a critique by Bound, Jaeger, and Baker (1995) it has been shown that if the correlation between instruments and endogenous regressors is small then standard asymptotic approximations of the distribution of IV estimators are unreliable both in small and large samples. From there, alternative asymptotic frameworks were developed to account for potentially weak identification or weak instruments, such as in the seminal paper by Staiger and Stock' (1997), so that robust tests and inference may still be performed, see e.g. , Anderson and Rubin (1949), Stock and Wright (2000), Moreira (2003), Kleibergen (2002, 2005), Andrews and Cheng (2012), Andrews (2016), and Andrews and Mikusheva (2016a,b). Important surveys on weak identification include Stock, Wright, and Yogo (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2005). These procedures are robust to instrument strength and rely on a parametric and often linear approximation of the first-stage equation. But using a linear approximation of the first stage leads to a loss of information and thus lowers IV strength, in a recent paper Dieterle and Snell (2016) highlights how in a variety of applications adding polynomials and cross-products of the instruments change the 2SLS estimates significantly. Consequently, in the context of weak instruments, it would be preferable to consider a non-linear first stage but as Jun and Pinkse (2012) have shown, using a nonparametric estimate of the first-stage conditional mean does not allow to obtain a valid confidence interval for β when instruments are weak. This negative results extends to nonparametric IV estimation procedures such as Newey and Powell (2003) or Darolles, Florens, and Renault (2011). In fact estimating the first stage conditional mean nonparametrically typically results in a weak-identification robust confidence interval which is wider than if the first stage was considered linear because the conditional mean is too flat and the number of IVs too large, see Dieterle and Snell (2016). For the above reasons Antoine and Lavergne (2022) came up with an inference procedure based on a test which leaves the first stage equation unspecified and unestimated while being robust to weak identification. Their test statistic uses the methodology of integrated conditional moments but is not pivotal hence its the distribution has to simulated. This makes their test hard to apply in practice when sample size becomes large ($n > 10,000$) as is most common in applied microeconomics

papers and in the application of this paper.

Consequently, I develop a pivotal test statistic for weak-identification robust inference in linear IV models with an unspecified first stage. This allows applied researchers to easily infer on the linear effect of endogenous variables on outcomes regardless of the nature of the relationship between the instruments and the endogenous variables. To create this test my approach resembles that of Antoine and Lavergne (2022): I combine the integrated conditional moment specification test of Bierens (1982) with the Lagrange multiplier test of Kleibergen (2005) (LM) to create KICM for Kleibergen integrated conditional moment, whereas they consider an integrated conditional moment version of Anderson and Rubin (1949) (AR) called ICM for integrated conditional moment and of the conditional likelihood ratio test of Moreira (2003) (CLR) called CICM for conditional integrated conditional moment. ICM and CICM are asymptotic tests, therefore they will perform well only in larger samples and yet they are also non-pivotal tests hence their critical values depend on the null hypothesis being tested and have to be simulated. This means that for large samples with possibly a few endogenous regressors it is very computationally costly to invert the ICM or CICM tests. On the contrary KICM is chi square with degrees of freedom equal to the number of endogenous regressors at the limit under the null with iid or heteroskedastic non-normal errors and a fixed number of instruments which makes inversion for inference relatively easy. In addition, it is known that the LM test is more robust to many and many weak instruments, see Hansen, Hausman, and Newey (2008), compared the AR and the CLR, this result carries on for their integrated conditional moment versions. These advantages shine in the simulations and application of this paper: With 4 instruments KICM has no size distortion compared ICM and CICM. The ICM and CICM tests cannot be inverted to infer on the effect of schooling on wages because sample size is above $100,000$.

In the second section of this paper I formally introduce the model, the existing tests, and motivate KICM. The third section is devoted to the derivation of the KICM test statistic from the null hypothesis and its implementation. The fourth establishes the validity and consistency of KICM for iid and heteroskedastic data. In the fifth section I perform an simulation exercise to assess KICM performances. In the sixth section I perform inference on the return to schooling on salary using the data from Angrist and Krueger (1991). I conclude in a seventh and final section. Proofs are in section A, B, C and D of the appendix, tables and plots from the simulations and the application are in section E of the appendix.

2 Framework

The objective is to infer on the effect β of l endogenous variables x_i on an outcome y_i by testing null hypotheses of the form $H_0 : \beta = \beta_0$ for some $\beta_0 \in \mathbb{R}^l$. Without loss of generality exogenous control variables are projected out'a la Frisch-Waugh consequently in the rest of the paper I consider the following structural equation

$$y_i = x_i'\beta + u_i \qquad \mathbb{E}(u_i|z_i) = 0 \qquad i = 1, \dots, n \tag{2.2}$$

which is augmented by a first-stage reduced form equation for x_i with $k > l$ exogenous instruments z_i

$$x_i = \Pi(z_i) + v_i \qquad \mathbb{E}(v_i|z_i) = 0 \qquad i = 1, \dots, n \tag{2.3}$$

z_i may also include some of the exogenous controls if one also suspects $\Pi(\cdot)$ to be non-linear in those. I denote by y, x, z and $\Pi(z)$ the stacked versions the versions of $(y_i, x_i', z_i', \Pi(z_i)')$ over the observations $i = 1, \dots, n$ so that y is of dimension $n \times 1$, x of dimension $n \times l$, z of dimension $n \times k$, and $\Pi(z)$ of dimension $n \times l$.

$\Pi(\cdot)$ may be "close to zero" so that z is weakly related to x. This weak instruments problem prevents consistent estimation of β and renders inference using standard Gaussian confidence interval invalid. Valid inference is still possible by inverting weak-identification robust tests but those may yield conservative confidence interval as they don't account for non-linearities in the first stage. In the next subsections I first briefly present the weak instruments problem, second I review the most popular methods for weak-identification robust inference, and third I motivate the use of KICM.

2.1 The weak instruments problem

Consider the setting described by (2.2) and (2.3) and assume for exposition that $(y_i, x_i, z_i)_{i=1}^n$ is iid and that $\Pi(\cdot)$ is linear and injective, ie $\Pi(z_i) = \Pi' z_i$ with Π a full rank $k \times l$ matrix. To estimate β one will use a k-class estimator such as 2SLS but when the instruments are weak as in Π is close to being singular the estimators mentioned above are biased and the traditional inference procedures become unreliable even in large samples. The literature has largely expanded upon these types of problems, see the surveys by Stock et al. (2002), Dufour (2003), Hahn and Hausman (2003), and Andrews and Stock (2005), and has coined different types of weak instruments asymptotics in order to model these problems. Because I consider

a finite number of instruments k, I follow the terminology of Andrews and Cheng (2012) and without loss of generality allow instruments to be very weak, weak, semi-strong and strong

$$\Pi \equiv \frac{1}{n^a} C \tag{2.4}$$

where C is a $k \times l$ full rank matrix and a is positive or infinite. a represents instruments' strength so that when $a = 0$ the instruments are deemed strong and β is strongly identified, when $a \in (0; 1/2)$ the instruments are deemed semi-strong and consequently β is semi-strongly identified, when $a = 1/2$ the instruments are deemed weak and thus β is weakly identified, and when $a > 1/2$ the instruments are deemed very weak and therefore β is very weakly identifiedMore specifically, as long as $a < +\infty$ the structural parameter β is point-identified however depending of the strength of the instruments β may not be consistently estimated which is why this terminology is used. When the instruments are weak or very weak $a \geqslant 1/2$ then k-class estimators such as 2SLS lose their consistency and their asymptotic normality.

Because consistent estimation is too difficult in case of weak instruments even with regularization, the literature has focused on providing inference robust to weak instruments by inverting tests. I introduce the most famous tests in the literature then show with a simple example that, because they do not take into account many non-linearties in the first stage, they decrease identification strength of β which is why KICM is needed.

2.2 Existing tests

Define

$$Y = \begin{pmatrix} y_1 & x_1' \\ y_1 & x_2' \\ \vdots & \vdots \\ y_n & x_n' \end{pmatrix}, \ \Omega_i \equiv \Omega(z_i) = \text{Var}(Y_i|z_i) = \begin{pmatrix} \text{Var}(y_i|z_i) & \text{Cov}(y_i, x_i|z_i) \\ \text{Cov}(x_i, y_i|z_i) & \text{Var}(x_i|z_i) \end{pmatrix} = \text{Var}(v_i'\beta + u_i \ \ v_i'|z_i)$$

$$b_0 = [1 \ \ -\beta_0']', \qquad A_0 = [\beta_0 \ I_l]'$$

Assuming the data is homoskedastic, ie $\Omega(z_i) = \Omega$, then one can also define

$$S \equiv S(\beta_0) = Y b_0 \left(b_0' \Omega b_0\right)^{-1/2}, \qquad T \equiv T(\beta_0) = Y \Omega^{-1} A_0 \left(A_0' \Omega^{-1} A_0\right)^{-1/2}$$

Notice that b_0 has dimension $(l+1)\times 1$ and A_0 has dimension $(l+1)\times l$ so T has dimension $n\times l$ whereas S has dimension $n\times 1$. Most notably, these notations imply that $Yb_0 = y - x\beta_0$ and that if the data is homoskedastic $\text{Var}(y_i - x_i'\beta_0|z_i) = b_0'\Omega b_0$.

Anderson and Rubin (1949) were the first to address the issue of inference with weak instruments under the assumption of homoskedacity and linearity in the first stage, without resorting to estimating the first stage correlation coefficient Π. The principle is the following: for different β_0, $y - x\beta_0$ is regressed on z and a test of joint significance of z is performed, then all the values of β_0 for which the test is not rejected form the confidence interval of β. To test $H_0 : \beta = \beta_0$ rewrite model (2.2)

$$Yb_0 = y - x\beta_0 = x(\beta - \beta_0) + u \equiv z\delta_0 + u_0$$

then the AR test statistic is the Wald statistic which tests $H_0 : \delta_0 = 0$

$$\text{AR} \equiv \text{AR}(\beta_0) = \frac{b_0'Y'P_zYb_0}{b_0'\Omega b_0} = S'P_zS$$

As z has rank k so does P_z which implies that under the null $H_0 : \beta = \beta_0$ and assuming linearity in the first stage and homoskedasticity $\text{AR} \xrightarrow{d} \chi^2_k$. This holds whether or not the errors are normals and if Ω is replaced by a consistent estimator such as $\frac{1}{n-k}Y'M_zY$.

Later came Kleibergen (2002) Lagrange Multiplier statistic

$$\text{LM} = S'P_{P_zT}S = S'P_zT(T'P_zT)^{-1}T'P_zS$$

derived from limited information maximum likelihood criterion. P_{P_zT} has rank l thus $\text{LM} \xrightarrow{d} \chi^2_l$ under the null, linearity of $\Pi(\cdot)$ and homoskedastic errors.

Moreira (2003) coined a conditional likelihood-ratio statistic

$$\text{CLR} = S'P_zS - \lambda_{min}\left(\begin{pmatrix} S'P_zS & S'P_zT \\ T'P_zS & T'P_zT \end{pmatrix}\right)$$

where $\lambda_{min}(\cdot)$ is the minimum eigenvalue. In general the asymptotic distribution of CLR has to be simulated and depends on β_0.

Note that $H_0 : \beta = \beta_0$ is equivalent to $\delta_0 = 0$ if and only if Π is non-singular hence if Π is singular or if the instruments are very weak $a > 1/2$, see (2.4), then the CI derived from any of these procedures will be of infinite length. Because the limiting distribution of LM does not depend on the number of instruments k but on the number of endogenous regressors l inference using LM yields confidence intervals with better coverage than with the AR and CLR if k is moderate or large. One downside of LM is that the confidence intervals for β may be the union of 2 or 3 intervals. When k is small it can be shown that CLR has better power than the other 2 and thus the confidence interval built from it has smaller length.

There exists heteroskedasticity robust and some auto correlation robust versions of the AR, LM and CLR tests, see e.g. Andrews, Moreira, and Stock (2004), Kleibergen (2007), Chernozhukov and Hansen (2008), Moreira and Moreira (2019) and Andrews and Mikusheva (2016a), tests which are similar in purpose but specific to other types of models such as the generalized empirical likelihood test of Guggenberger and Smith (2005) also exists. As noted by Dufour and Taamouti (2007) the AR test is "robust to misspecification" in the first stage unlike the LM and CLR tests: As long as there is one instrument left, if an instrument is not included in the first stage then the AR test will still be χ^2 distributed under the null thus it will have correct size. Tests which allow for non conservative inference on subvectors of β also exist, see Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), Guggenberger, Kleibergen, and Mavroeidis (2019).

2.3 Motivation

An important issue with these tests is that they only consider the linear relationship between the endogenous variables and the instruments hence non linearities remain in part undetected leading to a loss of power thus larger confidence intervals.

As an example consider the following scalar-IV model $l = k = 1$ with iid data homoskedastic data $(y_i, x_i, z_i)_{i=1}^n$ with Ω known and $z_i \sim \mathcal{N}(0, 1)$

$$y_i = x_i\beta + u_i, \qquad x_i = z_i^2 + v_i, \qquad \mathbb{E}(v_i|z_i) = \mathbb{E}(u_i|z_i) = 0$$

Then notice that the best linear projection of x_i on z_i denoted as $BLP(x_i|z_i)$ equals 0

$$BLP(x_i|z_i) = z_i'\mathbb{E}(z_iz_i')^{-1}\mathbb{E}(z_ix_i) = z_i'\mathbb{E}(z_iz_i')^{-1}\mathbb{E}(z_i^3) = 0$$

because $\mathbb{E}(z_i^3) = 0$. Similarly the projection y_i on z_i equals 0

$$BLP(y_i|z_i) = z_i'\mathbb{E}(z_iz_i')^{-1}\mathbb{E}(z_iy_i) = z_i'\mathbb{E}(z_iz_i')^{-1}\mathbb{E}(z_i^3\beta + z_iv_i\beta + z_iu_i) = 0$$

As a consequence, instruments are considered irrelevant by the AR, LM and CLR tests because all of them are quadratic functions of $Y'P_zY$, thus confidence intervals built from them are the whole real line. More precisely using the law of large numbers (LLN) and the central limit theorem (CLT) it can be shown that

$$Y'P_zY = \frac{1}{\sqrt{n}}Y'z\left(\frac{1}{n}z'z\right)^{-1}\frac{1}{\sqrt{n}}z'Y = O_{\mathbb{P}}(1)$$

where $O_{\mathbb{P}}(1)$ is the big O in probability notation for bounded in probability[1]. Thus AR cannot explode under the alternative $H_1 : \beta \neq \beta_0$ because it is bounded in probability, it has no power hence a confidence interval built from it will be very large.

In a more general case with the possibility of semi-strong and weak instruments, only considering linearities in the first stage as in the AR, LM and CLR can only exacerbate the issue of instrument weakness even in the best of cases while in the worst as in the example above it can make the instruments completely irrelevant. For this reason a new test which takes into account non-linearities in the first stage such as KICM is needed.

3 Building KICM

I derive the KICM statistic in two steps: First, I consider a conditional moment null hypothesis and prove that it is equivalent to an integrated conditional moment hypothesis as in Bierens (1982). Second, using the ICM statistic of Antoine and Lavergne (2022) as a criterion I build KICM which is a transformation of ICM's score. Then I present the feasible versions of KICM for both homoskedastic and heteroskedastic data. From now on and in the rest of the paper I consider the model characterized by (2.2) and (2.3) and unless specifically mentioned I do not assume that $\Pi(\cdot)$ is linear and that the data is iid.

3.1 From a conditional moment to an integrated conditional moment

Recall the model characterized by (2.2) and (2.3)

$$y_i = x_i'\beta + u_i \qquad \mathbb{E}(u_i|z_i) = 0 \tag{2.2}$$
$$x_i = \Pi(z_i) + v_i \qquad \mathbb{E}(v_i|z_i) = 0 \tag{2.3}$$

[1]Formally if $X = O_{\mathbb{P}}(1)$ then $\forall\varepsilon > 0 \exists M : \mathbb{P}(|X| > M) \leqslant \varepsilon$. If $X = o_{\mathbb{P}}(1)$ then X is degenerate in probability and $\forall\varepsilon > 0, \mathbb{P}(|X| > \varepsilon) \to 0$.

Then to test $H_0 : \beta = \beta_0$ the structural equation should be rewritten

$$y - x\beta_0 = \Pi(z)(\beta - \beta_0) + u_0, \qquad u_0 = v(\beta - \beta_0) + u$$

As a consequence, $H_0 : \beta = \beta_0$ implies that $H_0^1 : \mathbb{E}(y_i - x_i'\beta_0|z_i) = 0$ a.s which turns into an equivalence under specific conditions.

Using H_0^1 directly is not possible so instead I use the "Fourier" transformation from Bierens (1982) to obtain an equivalent many moments condition $H_0^2 : \mathbb{E}((y_i - x_i'\beta_0)e^{it'z_i}) = 0 \ \forall t \in \mathbb{R}^k$. One may interpret H_0^2 as the true error being 0 on average for any possible direction of the instruments or equivalently for any possible "additive combination" of the moments of the instruments. Indeed

$$\forall t \in \mathbb{R}^k \quad exp(it'z_i) = cos(t'z_i) + i\ sin(t'z_i) = exp(i\sum_{j=1}^{k} z_i^{t_k})$$

Alternatives to H_0^2 could be used such as a many moments condition with check functions, however using the complex exponential will allow to formulate the test in a simple matrix form and makes it pivotal.

The condition H_0^2 is equivalent to $H_0^3 : |\mathbb{E}((y_i - x_i'\beta_0)e^{it'z_i})|^2 = 0 \ \forall t \in \mathbb{R}^k$ where $|\cdot|$ denotes the modulus. Finally H_0^3 is equivalent to H_0^4, an integrated version of the many moments conditions over the t to only have 1 final moment so that the null $H_0 : \beta = \beta_0$ is equivalent to

$$H_0^4 : \int_{\mathbb{R}^k} |\mathbb{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0$$

where μ is a (finite) measure with support $\mathbb{R}^k$ which is positive almost everywhere to account for all the moments. These equivalences are summarized in the following proposition.

Proposition 3.1 *Assuming that (2.2) and (2.3) hold and that μ is a positive measure almost everywhere on $\mathbb{R}^k$ then*

$$\begin{aligned} H_0 : \beta = \beta_0 \Rightarrow H_0^1 : \mathbb{E}(y_i - x_i'\beta_0|z_i) = 0 \ a.s \ &\Leftrightarrow H_0^2 : \mathbb{E}((y_i - x_i'\beta_0)e^{it'z_i}) = 0 \ \forall t \in \mathbb{R}^k \\ &\Leftrightarrow H_0^3 : |\mathbb{E}((y_i - x_i'\beta_0)e^{it'z_i})|^2 = 0 \ \forall t \in \mathbb{R}^k \\ &\Leftrightarrow H_0^4 : \int_{\mathbb{R}^k} |\mathbb{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0 \end{aligned}$$

Moreover, if $l = 1$ *and* $\mathbb{P}(\Pi(z_i) = 0) = 0$ *then*

$$H_0 \Leftrightarrow H_0^1 \Leftrightarrow H_0^2 \Leftrightarrow H_0^3 \Leftrightarrow H_0^4$$

The proof is in A.1 of the appendix. It is then straightforward to build a test statistic for H_0 from H_0^4. Note that in general H_0^4 tests an implication of H_0 like the previously mentioned tests, when $\Pi(\cdot)$ is small the test has low power and the confidence interval built from it is large.

3.2 From ICM to KICM

To test $H_0 : \beta = \beta_0$ an empirical counterpart of $H_0^4 : \int_{\mathbb{R}^k} |\mathbb{E}((y_i - x_i'\beta_0)e^{is'z_i})|^2 d\mu(s) = 0$ is taken, then multiplying by n and standardizing allows the CLT to apply to the integrand, this is the ICM statistic of Antoine and Lavergne (2022) which writes

$$\text{ICM} \equiv \text{ICM}(\beta_0) = \int_{\mathbb{R}^k} |n^{-1/2} \sum_{i=1}^{n} \frac{y_i - x_i'\beta_0}{\text{Var}(y_i - x_i'\beta_0|z_i)^{1/2}} e^{is'z_i}|^2 \mu(s)$$

ICM can actually be written as a function of S, let W be a $n \times n$ matrix with elements $W_{ij} = n^{-1} w\left(z_i - z_j\right)$ such that

$$w(z) = \int_{\mathbb{R}^k} e^{is'z} d\mu(s) \tag{3.5}$$

The condition for μ to have support $\mathbb{R}^k$ translates into the restriction that $w(\cdot)$ should have a Fourier transform which is strictly positive almost everywhere, or if the support of the instruments z is bounded, that its Fourier transform is well-defined in a neighborhood of 0, see Theorem 1 in Bierens (1982). The choice of $w(\cdot)$ thus includes products of densities such as triangular, normal, or logistic, see Johnson, Kotz, and Balakrishnan (1995), Student, including Cauchy, see Dreier and Kotz (2002), or Laplace. Using properties of the modulus it is simple to show that ICM $= S'WS$.

Thus ICM ressembles AR $= S'P_zS$ but W is not a projection matrix hence ICM is not pivotal asymptotically and its distribution depends on β_0. Similarly from Antoine and Lavergne (2022) CICM is the integrated conditional moment equivalent of the CLR of Moreira (2003) and is not pivotal asymptotically. To derive the KICM test statistic which is pivotal asymptotically I use ICM as a criterion function and derive its score which I then standardize: Taking

(y, x, z) as deterministic notice that the convex function

$$\beta \mapsto \frac{b'Y'WYb}{b'\Omega b}, \quad b = (1 \ -\beta')'$$

is minimized uniquely at $\beta = \beta_0$ under the null hypothesis. Thus taking the first order condition at β_0 yields

$$\begin{aligned}
\frac{\partial}{\partial\beta}b_0 \times \left(\frac{Y'WYb_0}{b_0'\Omega b_0} - \frac{\Omega b_0 b_0' Y'WYb_0}{(b_0'\Omega b_0)^2}\right) = 0 \ &\Leftrightarrow \frac{Y'WYb_0}{b_0'\Omega b_0} - \frac{\Omega b_0 b_0' Y'WYb}{(b_0'\Omega b_0)^2} = 0 \\
&\Leftrightarrow \frac{(A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}Y'WYb_0}{(b_0'\Omega b_0)^{1/2}} = 0 \\
&\Leftrightarrow T'WS = 0
\end{aligned}$$

where the second line is obtained by multiplying by $(A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}$ and using the fact that $A_0'b_0 = 0_l$. I prove later that $\mathbb{E}(S'WT|z) = 0$ when (y, x, z) is random. Finally KICM is a quadratic version of the score $S'WT$ standardized with respect to WT

$$\text{KICM} = S'WT(T'W^2T)^{-1}T'WS = S'P_{WT}S$$

where P_{WT} is the orthogonal projection matrix on WT which gives the statistic its chi square l degrees of freedom asymptotic distribution.

3.3 KICM in practice

In practice in order to use KICM properly several elements are still needed:

First, I greatly simplify the choice of $w(\cdot)$ by imposing that its Fourier transform is a real symmetric density which is strictly positive almost everywhere, or around 0 if z has bounded support, and that the L_2 norm of $w(\cdot)$ equals 1. As a consequence, $w(\cdot)$ is also a symmetric real bounded density, thus possible choices for $w(\cdot)$ are Triangular, Logistic, Cauchy or Laplace distribution densities (see Johnson et al. (1995)). Imposing that the Fourier transform is a centered density cleverly prevents having too many instruments, puts more weights on lower moments of z and make sure no cardinal direction of moments of z is favored. On the other hand making sure that the squared norm of $w(\cdot)$ equals 1 ensures that the elements of W do not scale with sample size.

Second, $\Omega(z_i) = \text{Var}(Y_i|z_i)$ must be estimated consistently. Assuming it is linear in z then simply using the parametric estimator $\hat{\Omega} = \frac{1}{n}Y'M_zY$ is a good idea. If this assumption is too strong one may use a semi-parametric or non-parametric estimator, e.g. from Seifert, Gasser,

and Wolf (1993) or from Yin, Geng, Li, and Wang (2010). I use the later in the simulations and application. It writes

$$\hat{\Omega}(z) = \frac{\frac{1}{nh}\sum_{i=1}^{n}(Y_i - \bar{Y}(z))(Y_i - \bar{Y}(z))' K((z_i - z)/h)}{\frac{1}{nh}\sum_{i=1}^{n} K((z_i - z)/h)}, \; \bar{Y}(z) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K((z_i - z)/h)}{\frac{1}{nh}\sum_{i=1}^{n} K((z_i - z)/h)}$$

with the bandwidth h chosen properly to allow convergence. If data is homoskedastic the estimator is the following average $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}(z_i)$.

Third, the instruments should be standardized. A desirable property of weak-identification robust tests is that of invariance to orthogonal transformations of the instruments which allows the tests to be invariant to instruments' scale (see Andrews and Stock (2007)). KICM cannot satisfy this property however by standardizing the instruments a priori the same effect can be obtained. Additionally in the literature on nonparametric estimation via Kernels, regressors are standardized through the bandwidth (see Li and Racine (2006)) and here $w(\cdot)$ has the role of a Kernel function.

Feasible tests Based on the above the feasible KICM statistic for homoskedastic data $\text{Var}(Y_i|z_i) = \Omega$ writes

$$KICM_f = S_f' P_{WT_f} S_f, \quad S_f = Y b_0 (b_0' \hat{\Omega} b_0)^{-1/2}, \quad T_f = Y \hat{\Omega}^{-1} A_0 (A_0' \hat{\Omega}^{-1} A_0)^{-1/2}$$

where W has elements $W_{ij} = \frac{1}{n} w(z_i - z_j)$ with $w(\cdot)$ a density satisfying the aforementioned conditions and where $\hat{\Omega}$ is a consistent estimator of Ω.

In case of heteroskedastic data $\text{Var}(Y_i|z_i) = \Omega(z_i) = \Omega_i$ I first define the heteroskedasticity robust version of KICM

$$\text{KICM}_h = S_h' P_{WT_h} S_h, \quad \forall i \;\; S_{ih} = Y_i' b_0 (b_0' \Omega(z_i) b_0)^{-1/2}, \quad T_{ih}' = Y_i' \Omega(z_i)^{-1} A_0 (A_0' \Omega(z_i)^{-1} A_0)^{-1/2}$$

with S_h and T_h the stacked versions of S_{ih} and T_{ih} respectively. This change allows S_{ih} and T_{ih} to be properly standardized in the heteroskedastic case. Then the feasible version of the heteroskedasticity robust KICM statistic writes

$$\text{KICM}_{hf} = S_{hf}' P_{WT_{hf}} S_{hf}, \quad \forall i \;\; S_{ihf} = Y_i' b_0 (b_0' \hat{\Omega}(z_i) b_0)^{-1/2}, \quad T_{ihf}' = Y_i' \hat{\Omega}(z_i)^{-1} A_0 (A_0' \hat{\Omega}(z_i)^{-1} A_0)^{-1/2}$$

where $\hat{\Omega}(\cdot)$ is a consistent estimator of $\Omega(\cdot)$.

Weak-identification robust inference With a feasible KICM test statistic in hand it is now possible to infer on β regardless of its degree of identification. To do so the econometrician has to invert the KICM test. First, they need to select a nominal coverage $1-\alpha$ for the confidence interval and a grid over $\mathbb{R}^l$ from which they will test different values of β_0. Then the econometrician must compute the KICM feasible statistic for all values of β_0 over the grid. Finally all values of β_0 for which the statistic is above the $1-\alpha$ quantile of a chi square with l degrees of freedom will constitute the $1-\alpha$ confidence interval of β.

The next section is devoted to formal results on the distibution and the asymptotic behavior of KICM.

4 KICM validity and consistency

In this section, I first introduce assumptions necessary for the asymptotic theory then I prove that KICM is chi square distributed under normality of the errors under the null, and lastly I prove validity and consistency of KICM without normality of the errors. The theoretical coverage probabilities of the confidence interval built from KICM are direct implications of the propositions and theorem introduced in this section which is why they are omitted.

4.1 Assumptions

I first assume that the data is either iid or heteroskedastic in the sense that errors' variance are functions of the instruments. Then to obtain asymptotic results and a consistent estimator of the conditional variance $\Omega(\cdot)$ I require Y_i to have strictly more than a second conditional moment which is bounded in order to use Berry-Esseen inequalities to prove convergence.

Assumption A
(i) Observations $(y_i, x_i', z_i')_{i=1}^n$ *are independent and identically distributed*
(ii) Observations $(y_i, x_i', z_i')_{i=1}^n$ *are independent with* $(z_i')_{i=1}^n$ *also identically distributed*
(iii) $\exists \delta > 0, M > 0 : \mathbb{E}(||Y_i||^{2+\delta}|z_i) < M$

Second, I make assumptions on the parameters. I assume the unique existence of a structural parameter of interest β and of some reduced form parameter $\Pi(\cdot)$. Then in order to model strong, weak and very weak identification I allow $\Pi(\cdot)$ to depend on n in two ways: Either $\Pi(\cdot) = n^{-a}C(\cdot)$ where $C(\cdot)$ is a function which does not depend on n and a represents

the degree of identification of β or equivalently the degree of weakness of the instruments. The coefficient a is just a theoretical tool to study size and power when parameters have different identification strength, it is unknown in practice. Then β is strongly identified or equivalently instruments are strong when $a = 0$, β is semi-strongly identified and the instruments are semi-strong when $0 < a < 1/2$, β is weakly identified and the instruments are said to be weak in the sense of Staiger and Stock' (1997) when $a = 1/2$, β is very weakly identified and instruments are very weak when $a > 1/2$, and when $a = \infty$ instruments are irrelevant and β is not identified at all. Either $\Pi(\cdot) = N_C^{-1}C(\cdot)$ where N_C is a $l \times l$ diagonal matrix with entries which correspond to the degree of identification of each element in the vector β or equivalently the degree of weakness of the instruments with regards to each element of β

$$N_C = \begin{pmatrix} n^{a_1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n^{a_l} \end{pmatrix}, \quad (a_1, \dots, a_l) \in \bar{\mathbb{R}}_+^l$$

where a_j represents the degree of identification of β_j.Thus if $a_j = 0$ instruments are strong for j and β_j is strongly identified, if $0 < a_j < 1/2$ then β_j is semi-strongly identified, etc... With either of those assumptions β and its elements can be strongly identified, weakly identified or not identified at all if $a = \infty$. Lastly, I assume $C(z_i)$ to have a strictly positive and finite second moment.

Assumption B
(i) There exists a unique β and some $\Pi(\cdot)$ such that (2.2) and (2.3) hold
(ii) $\Pi(\cdot) = n^{-a}C(\cdot)$ where $a \in \bar{\mathbb{R}}_+$
(iii) $\Pi(\cdot) = N_C^{-1}C(\cdot)$ where N_C is a diagonal matrix with entries $(n^{a_j})_{j=1}^l$ with $a_j \in \bar{\mathbb{R}}_+$
(iv) $C(\cdot)$ does not depend on n and $0 < \mathbb{E}(C(z_i)C(z_i)') < +\infty$

Next I impose conditions on $w(\cdot)$ so that by Bochner's Theorem $\mu(\cdot)$ is finite and strictly positive almost everywhere. Furthermore, I assume that W is positive definite which holds in practice but cannot be formally proven without imposing more conditions on $w(\cdot)$ and z.

Assumption C
(i) $w(\cdot)$ has a Fourier transform which is strictly positive almost everywhere, or which is strictly positive in a neighborhood of 0 if the support of z_i is bounded
(ii) W is positive definite almost surely for any n

Lastly, I impose conditions on the conditional covariance estimator $\hat{\Omega}$. In the homoskedastic case a simple consistent estimator for any possible DGP is needed.

In the heteroskedastic case proving the validity and consistency of KICM involves looking at random processes of Ω and t such as $(t, \Omega) \mapsto Y_i' \Omega^{-1} cos(t' z_i)$. These processes must be sufficiently smooth in Ω which is why I restrict the covariances to the class $\mathcal{O}$. Each Ω in the class $\mathcal{O}$ should be "uniformly bounded" using their minimal and maximal eigenvalues. Functions in that class also have to be smooth enough so that the class is not "too large", for any DGP $\mathcal{O}$ has a finite covering number[2] denoted $N(\varepsilon, \mathcal{O}, L_2(\mathbb{P}))$ which should not to explode when ε gets bigger for any DGP. This assumption is necessary in order to obtain asymptotic equicontinuity uniform of the random processes involving Ω so that the difference between the feasible version KICM and KICM vanishes in the heteroskedastic case, see Vaart and Wellner (2000), Kosorok (2008) and the proof of Theorem 4.3 in appendix B for more details. In addition, estimator $\hat{\Omega}(\cdot)$ also needs to converge uniformly towards $\Omega(\cdot)$ in the L_2 sense and to belong uniformly to $\mathcal{O}$ almost surely at the limit. In the literature on convergence of nonparametric statistics with nuisance parameters this type of condition is common, see Andrews (1995), these are also necessary conditions for the difference between feasible KICM and KICM to vanish uniformly at the limit.

Assumption D

Let $\mathcal{P}$ denote the set of all distributions which satisfy assumptions A(i), A(ii) or A(iii), B(i), B(ii) or B(iii), B(iv), C(i) and C(ii)

(i) $\forall \mathbb{P} \in \mathcal{P}\ \hat{\Omega} \xrightarrow{P} \Omega$

(ii) Define $\mathcal{O}$ such that $\forall \Omega(\cdot) \in \mathcal{O}$

$$0 < \underline{\lambda} = \inf_{s \in \mathbb{R}^l, \Omega \in \mathcal{O}} \{\lambda(\Omega(s))\} < \sup_{s \in \mathbb{R}^l, \Omega \in \mathcal{O}} \{\lambda(\Omega(s))\} = \bar{\lambda} < +\infty$$

$$\forall \mathbb{P} \in \mathcal{P},\ \forall \varepsilon > 0,\ N(\varepsilon, \mathcal{O}, L_2(\mathbb{P})) < K' e^{-K\varepsilon}$$

for some $K' > 0$, $K < 2$

(iii) $\sup_{P \in \mathcal{P}} ||\hat{\Omega}(\cdot) - \Omega(\cdot)||_{L_2(\mathbb{P})} \to 0,\ \sup_{P \in \mathcal{P}} \mathbb{P}(\hat{\Omega}(\cdot) \in \mathcal{O}) \to 1$

4.2 Normal Errors

For exposition I first consider conditionally normal errors or equivalently that $Y_i | z_i \sim \mathcal{N}(\mathbb{E}(Y_i | z_i), \Omega)$. Under this assumption it is relatively simple to show that under the null KICM $\sim \chi^2_l$ if the

[2] A ball $L_2(\mathbb{P})$ of size $\varepsilon > 0$ centered in $g \in L_2(\mathbb{P})$ writes $\{f \in L_2(\mathbb{P}) : \int ||f - g||_2^2 d\mathbb{P} < \varepsilon\}$, then the covering number of $\mathcal{O}$ for DGP P with interval of size ε is the minimal number of ε-balls necessary to cover all of $\mathcal{O}$

data is homoskedastic, and $\text{KICM}_h \sim \chi^2_l$ if the data is heteroskedastic.

In the homoskedastic case $\text{Var}(Y_i|z_i) = \Omega$ it can be shown that WT has rank l and therefore that the orthogonal projection matrix P_{WT} also has rank l. In turn using the fact that S and T are orthogonal this implies that KICM is χ^2_l distributed conditionally on (z, T) which implies that KICM is χ^2_l distributed. This result is summarized in the following proposition.

Proposition 4.1
For any $\mathbb{P} \in \mathcal{P}$ such that $H_0 : \beta = \beta_0$, assumptions A(i), B(i), B(iii), B(iv), C(i) and C(ii) hold and assuming that $Y_i|z_i \sim \mathcal{N}(\mathbb{E}(Y_i|z_i), \Omega)$ holds then $\text{KICM} \sim \chi^2_l$

The proof is in A.2 of the appendix. As a direct corollary, the feasible statistic under normal errors KICM_f is also asymptotically χ^2_l when assumptions A(iii) and D(i) are added. The reason is that, under the null, KICM_f is bounded in probability, and that S_{if} and T_{if} have covariance identity at the limit and therefore S_f and T_f are orthogonal at the limit. This result is an implication of the continuous mapping theorem (CMT) so its formal proof is omitted.

In the heteroskedastic case $\text{Var}(Y_i|z_i) = \Omega(z_i)$ the heteroskedasticity robust KICM statistic KICM_h also follows a chi square distribution with l degrees of freedom

Proposition 4.2
Given Ω, for any $\mathbb{P} \in \mathcal{P}$ such that $H_0 : \beta = \beta_0$, assumptions A(i), B(ii), B(iii), B(iv), C(i) and C(ii) hold and assuming that $Y_i|z_i \sim \mathcal{N}(\mathbb{E}(Y_i|z_i), \Omega(z_i))$ holds then $\text{KICM}_h \sim \chi^2_l$

The proof is very similar to that of Proposition 4.1 and is omitted. Once again the key argument is that (S_{ih}, T_{ih}) has variance identity conditionally on z_i. The feasible statistic KICM_{hf} can also be proven to be χ^2_l asymptotically taking $\Omega(\cdot)$ as given. The formal proof of this result is also omitted.

4.3 Non-normal Errors

With non-normal errors I use empirical process theory to prove that the heteroskedasticity robust feasible KICM test is uniformly valid and uniformly consistent. In the heteroskedastic case $\Omega(z_i)$ is random and could be unbounded thus it is very desirable that KICM is valid and consistent for all possible data generating processes and not specific ones. Note that the following results also hold under homoskedastic data and / or normal data.

Theorem 4.3 shows that the feasible heteroskedasticity robust KICM test is uniformly valid regardless of instruments strength. Indeed no statement is made about $(a_j)_{j=1}^l$ thus by inverting this test one automatically obtains a confidence interval with at least nominal coverage. Of course in practice this interval may be large, especially if instruments are weak.

Theorem 4.3 (Uniform Validity of KICM)
Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom. Then, under the null $H_0 : \beta = \beta_0$ and assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii) and D(iii):

$$\lim_{n\to\infty} \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta=\beta_0} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha$$

The size of KICM *is lesser or equal than its nominal size asymptotically.*

The proof of Theorem 4.3 is in B of the appendix.

Regarding the power of the KICM test it is determined by the identification strength of the elements of β which differs from β_0, in addition it is the element which is best identified which will drive power. Intuitively, to reject H_0 the part of β which is different from β_0 must be at least semi-strongly identified. Mmore explicitly if for element j and j' $\beta_j = \beta_{j0}$ and $\beta_{j'} \neq \beta_{j'0}$, then element j can never contribute to rejecting the null whereas j' will contribute if $a_{j'} \leqslant 1/2$. Thus if B(ii) is assumed the test is consistent if $a < 1/2$ but if B(iii) is assumed then the test is consistent if $min\{a_j : \beta_{j0} \neq \beta_j\} < 1/2$. For simplicity, Corollary 4.4 presents the asymptotic uniform power properties of the feasible heteroskedasticity robust KICM test under assumption B(ii), ie $\Pi = n^{-a}C(\cdot)$. The power properties of the KICM test under assumption B(iii), ie $\Pi(\cdot) = N_C^{-1}C(\cdot)$, are presented and discussed in appendix D.

Corollary 4.4 (Uniform consistency of KICM)
Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom. Then under assumptions A(ii), A(iii), B(i), B(ii), C(i), C(ii), D(ii), and D(iii),

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, a<1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) = 1;$
  The test is consistent when the instruments are at least semi-strong.

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, a=1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) \in [\alpha; 1);$
  The test has more than trivial power when the instruments are weak.

- $\lim_{n\to\infty} \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, a>1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha;$
  The test has trivial power when the instruments are very weak.

The proof of Corollary 4.4 is in C of the appendix. Next, I study the empirical performances of KICM.

5 Simulations

Setting I perform simulations in order to evaluate the empirical performances (size, power, confidence interval length) of the KICM test in small samples in case of strong, semi-strong or weak instruments, for 4 different first stages, and in case of homoskedastic or heteroskedastic data. The specification in the simulations is the following: I assume that there is one regressor $l = 1$ and either one or two instruments $k \in \{1, 2\}$ in (2.2) and (2.3), the true β is 0, the instrument z_i are standard normal thus centered, uncorrelated, and with a symmetric distribution, sample size is either 100 or 400. Thus there are 24 possible setups:

- 4 possible first stages: linear; non-linear; polar polynomial; semi-polar polynomial

$$\Pi_1(z) = \frac{z_1}{n^a}, \qquad \Pi_2(z) = \frac{1}{n^a}\frac{z_1 + z_2 + z_1 z_2 + z_1^2 + z_2^2 + z_1^2 z_2^2 - 3}{\sqrt{26}},$$
$$\Pi_3(z) = \frac{1}{n^a}\frac{z_1^2 - 1}{\sqrt{3}}, \qquad \Pi_4(z) = \frac{1}{n^a}\frac{z_1 + z_2^2 - 1}{\sqrt{4}}$$

- 3 instrument strengths: strong $a = 0$; semi-strong $a = 1/4$; weak $a = 1/2$
- 2 data types: homoskedastic; heteroskedastic

$$\Omega = \begin{pmatrix} 1 & 0.81 \\ 0.81 & 1 \end{pmatrix}, \qquad \Omega(z) = \frac{1 + z_1^2}{2}\begin{pmatrix} 1 & 0.81 \\ 0.81 & 1 \end{pmatrix}$$

Keeping instrument strength constant the data $(y_i, x_i, z_i)_{i=1}^n$ has the same mean and variance whatever the setup. Note that for this reason, controlling instruments strength a is equivalent to controlling the value of the (nonlinear) concentration parameter $\mu^2 = \frac{\Pi(z)'\Pi(z)}{\sqrt{\text{Var}(v_i)}}$ as in papers which define instruments' strength by the value of the concentration parameter, see Stock and Yogo (2005), Staiger and Stock' (1997). In these simulations if $a = 1/2$ then $\mu^2 \approx 1$, if $a = 1/4$ then $\mu^2 \approx \sqrt{n}$, if $a = 0$ then $\mu^2 \approx n$.

Competing methods I consider 6 competing procedures for building confidence sets: the AR, the LM, the CLR, the ICM, the CICM and the Wald, the later is simply the confidence

set built using the 2SLS estimator and using its traditional confidence interval based on the t-test Gaussian asymptotics. Note that if the first stage is polar polynomial as specified above then the best linear projection of y_i and x_i on z_i is 0, and that if the first stage is semi-polar polynomial then the best linear projection of y_i and x_i on z_{2i} is 0. So even in case of strong instruments I expect the AR, LM and CLR tests to perform very badly in terms of power compared to the KICM (and ICM and CICM). The number of simulations required to build the CI of the CLR, ICM and CICM vary between $m = 200$ and $m = 500$. Finally to create comparable heteroskedasticity robust versions of the AR, LM, CLR, ICM, KICM and CICM I use the nonparametric estimator of $\Omega(\cdot)$ of Yin et al. (2010), and consider the Eicker-White estimator of the covariance matrix of the 2SLS estimator for the Wald test.

Empirical size First, the coverage of the confidence intervals built with KICM is of special interest, it is the probability that the true β, which equals 0 in this setting, is in said interval (which is random). The empirical coverage is equal to the empirical size when tests are inverted. Hence, when comparing sizes the best procedure is the one for which the empirical size is closest to nominal size which I set to 10%. I report the empirical sizes of the AR, LM, CLR, ICM KICM, CICM and Wald test for the different settups in table 1 in the weak instruments case, in table 2 in the semi-strong instruments case, and in table 3 in the strong instruments case constructed over 5000 simulations in appendix E.1.1.

From the tables, the AR, LM, CLR, ICM, KICM and CICM are all robust to weak instruments thus in terms of size there is little difference between them when the strength of instruments changes. This is not the case for the Wald test built from 2SLS, it is not robust to weak instruments nor non-linearities, in fact even with semi-strong instruments it is very oversized. More precisely in all settings it seems that the empirical size of KICM is closer to nominal size 10% than both ICM and CICM which require a larger number of simulations in order to be competitive, especially in the linear case $\Pi_1(\cdot)$. In addition, KICM has better size than the AR, LM and CLR in all settings expect in the linear one which is within expectations.

Power curves Second, to assess how KICM rejects wrong values of β_0 I plot its power curve and the power curves of other competing tests as is done in most of the literature on testing. A power curve is drawn by measuring the empirical probability of rejecting the null $H_0 : \beta = \beta_0$ for many different β_0 in a grid. This implies that at the grid point $\beta_0 = \beta$ it is the empirical size of the test that is computed. Power curves are a useful tool as they tell if one test will reject false values of β more than another, consequently if a test's power curve dominates

another's then its confidence intervals will be systematically tighter than its competitor's.

In appendix E.1.2 below are the power curves of the AR, LM, CLR, ICM, KICM, CICM and Wald test built from 5000 replications for a test of nominal size 10% for the linear, non-linear and polar polynomial first stages. Sample size is $n = 400$ and I used $m = 500$ simulations in order to compute the critical values of the CLR, ICM and CICM tests. Figure 1, figure 2, figure 3 are the power curves of the 7 tests for strong, semi-strong and weak instruments respectively with homoskedastic data. Figure 4, figure 5, figure 6 are the power curves for strong, semi-strong and weak instruments respectively with heteroskedastic data.

Several remarks should be made: Notice that the curves are almost similar whether data is homoskedastic or heteroskedastic. Next in the polar case and non linear case note that the tests which are non-robust to non-linearities (AR, LM CLR and Wald) experience a large loss of power even when instruments are strong. All tests have trivial power when instruments are weak which again is as expected. Additionally KICM has power overall similar to ICM and CICM except on one side, this is due to the fact that inverting the KICM test is equivalent to solving a quartic inequality, LM has similar properties, see Mikusheva (2010). Thus when choosing between KICM, ICM and CICM there may be a tradeoff between coverage and power. KICM always seem to have better coverage but CICM seem to have better power in some cases.

Average p-value curves To have an idea of the "average length" of the CI interval for each test procedure, I define an "average" confidence interval built by inverting a test over many simulations. I could consider taking the average bounds of the confidences intervals built over many simulations, however bounds may not exists when instruments are weak, so instead I find the average 90% coverage confidence interval by using average p-value curves: For any candidate β_0 for any test for any setting, I check if the average p-value when testing $H_0 : \beta = \beta_0$ is above 10%. Then all the β_0 for which it is true will constitute the average 90% confidence interval.

Figure 7, 8 and 9 in appendix E.1.3 are plots of the average p-value curves for strong, semi-strong and weak instruments respectively, with heteroskedastic data, $n = 400$, $m = 500$.

The sets are the whole real line in case of weak instruments for all tests except the Wald test which gives a finite interval which has very low coverage. The tests which are non robust to non-linearities (AR, LM and CLR) have higher average p-values hence the average confidence intervals built from them are much large than the sets built from ICM, KICM and CICM in case of non-linear first stage, and infinite in case of polar polynomial first stage unlike sets

built from ICM, KICM and CICM. There is little to differentiate the sets built with KICM from the ones built from ICM and CICM in the strong or weak instruments case. In case of semi-strong instruments however the set built from KICM is the union of 2 finites sets in the linear and non-linear first stage case, one big set which is common the ICM and CICM and one much smaller set. Again there seems to be a tradeoff between having the right coverage and higher power when choosing between KICM and ICM and CICM.

Empirical size with a higher number of instruments Finally I consider a first stage with a higher number of instruments. It is well known that the LM test fares much better when k starts to grpw compared to both the AR and CLR in the linear case. The same holds true for KICM compared to ICM and CICM. I consider the linear first stage

$$\Pi_5(z) = \frac{1}{n^a}\frac{z_1 + z_2 + z_3 + z_4}{\sqrt{4}}$$

Then the empirical coverage of the 7 tests over 5000 simulations for homoskedastic data, sample size $n = 100$ and $m = 200$ simulations of the distributions of CLR, ICM and CICM are in table 4.

Clearly all the tests are oversized except the LM and KICM, hence a confidence interval built from will have a lower coverage than the nominal 90% for weak, semi-strong or strong instruments.

6 Application: Returns to schooling

In this final section I provide inference via KICM for the causal effect of the number of years of schooling on the logarithm of wage using quarters of birth as instruments using the data from Angrist and Krueger (1991). The authors estimate the causal effect of the number of years spent in school on wage by using the exogenous variation of schooling due to difference in quarters of births: Teenagers born early in the year leave school earlier because they reach the age at which they can work earlier, this creates a difference in the total number of years of schooling between children born early in the year and children born later. In addition, the authors try different specifications by interacting these instruments with time and location dummies in order to increase the fit of the first stage with the belief that it increases the strength of the instruments. Data is from the US where they have access to different cohorts and I focus on cohort 20-29 with 247,199 observations in the sample.

This paper is well-known for the fact that the instruments used are quite weak[3], across all specifications and all cohorts the F-statistic vary between 1 and 15. Because of the large sample size ICM and CICM cannot be used. At the same time considering that the first stage is non-linear allows quarter of births to act as types, the first stage is equal to a different non-linear function of the exogenous regressors for each type. This degree of flexibility is much larger compared to a linear first stage with only a few interactions being considered. Thus one can expect the confidence intervals built from KICM to be small compared to the competition if there is a sufficient number of covariates.

Formally I consider the model in table IV of Angrist and Krueger (1991) which focuses on cohort 20-29

$$\begin{aligned} log(wage)_i &= \beta\ schooling_i + FE_y + FE_r + x_i'\gamma + u_i \\ schooling_i &= \sum_{j,t} \alpha_{jt}\ 1_{QB_i=j,YB=i} + FE_y + FE_r + x_i'\zeta + v_i \end{aligned}$$

where QB_i is the quarter of birth, YB_i is the year of birth, FE_y are year of birth fixed effects, FE_r are region of residence fixed effects, and x_i some covariates. Using KICM I allow the first stage to be completely non-linear in the instruments and covariates thus I consider

$$schooling_i = \sum_{j,t,r} 1_{QB_i=j,YB_i=t,RR_i=r} \Pi_{jtr}(x_i) + v_i$$

where RR_i is the region of residence, and $\Pi_{jtr}(\cdot)$ is a non-linear function of x_i specific to the triple quarter of birth j, year of birth t, and region of residence r. In other words when using KICM I implicitly allow for a first stage with a different non-linear function for every possible combination of quarter of birth, year of birth and region of residence. This is much more flexible than a linear first stage specification.

In table 5 of appendix E.2 below I provide estimates of the 90% coverage heteroskedasticity robust confidence interval of β by inverting AR, LM, CLR, and KICM and for reference I also provide the OLS, 2SLS, LIML and Fuller estimates and their t-test Gaussian based confidence intervals for 4 different specifications.

In the simplest specification (1) there are no covariates only year fixed effects therefore KICM considers the same first stage as the other tests. At the same time without covariates,

[3]see Stock and Yogo (2005) for a comprehensive look at thresholds which determine if instruments are weak

schooling is very likely to still be endogenous even after being projected on the instruments. Assuming exogeneity however, observe that even in the simplest specification the KICM confidence set for returns to schooling is positive and has smaller length than the set built with the AR test, the CLR set however is significantly smaller. For specification (2) and (3) which add other covariates all 4 test procedures which are robust to weak instruments give the whole real line as the confidence interval. Finally in specification (4) which also includes region of residence fixed effects, while other tests give the whole real line as confidence sets, the KICM confidence interval is small and positive. In figure 10 of appendix E.2 is the p-value curve for KICM and Wald tests built using the different estimators over a grid of potential null $H_0 : \beta = \beta_0$ in specification (4). This result is not so surprising as KICM becomes more powerful with more covariates as it considers all non-linearities including interactions. This set is also very different from the OLS and 2SLS set but is included in the LIML and Fuller set which is not far-fetched because LIML and Fuller are known to perform better than 2SLS when there are weak possibly many instruments.

These results imply that estimates of the returns to schooling may not be as small as OLS and 2SLS and not as large as Fuller and LIML have indicated until now, from specification (4) an increase of 1 year of schooling yields an increase in wages between 13.8% and 24%.

7 Conclusion

On the one hand in the current literature on weak instruments, most inference procedures do not take into account non-linearities or interactions in the first stage. This leads to an important loss of relevance, or a total loss of relevance of the instruments, both in simulations and in applications. On the other hand estimating the first stage non-linearly is difficult and leads to a situation of having too many weak instruments, see Dieterle and Snell (2016). Thus like ICM and CICM from Antoine and Lavergne (2022), KICM relies on an integrated conditional moment in order to consider the non-linearities present in the first stage.

KICM has some advantages over the ICM and CICM. First, its size is closer to nominal size in practice hence sets built with KICM have coverages which are closer to nominal coverage compared to ones built with ICM or CICM. Second, it is more robust to many instruments than ICM and CICM just like how the LM is more robust to many instruments than the AR and CLR. Third, it is pivotal hence it does not require simulations in order to obtain its asymptotic distribution, ICM and CICM require simulations and are thus unimplementable when samples get large as in the application of this paper. This implies that in larger samples

or with more instruments KICM is very simple to implement and reliable. All in all KICM is an off-the shelf procedure to easily compute confidence sets which are robust to weak instruments and which consider non-linearities in the first stage, regardless of sample size or the number of instruments, normality or non-normality of the data, homoskedasticity or heteroskedasticity of the data.

Consequently, linear IV models with a single instrument and few covariates and / or fixed effects should expect significant improvements in the quality of confidence sets at no cost when using KICM compared to alternative procedures as it will automatically consider both interactions and non-linearities in the first stage while maintaining pivotality of the test. The pivotality of KICM is very useful in applied microeconomic settings with large samples, but this property is lost in practice when data is clustered or auto-correlated without imposing a lot of structure, this requires more investigation. In addition, there exists other pivotal tests for weak-identification robust inference which consider a non-linear first stage and these could have better power properties than ICM, CICM or KICM. As in the work of Moreira and Ridder (2017) or Andrews and Mikusheva (2016a), it may be possible to prove that ICM, CICM and KICM are or are not optimal.

Bibliography

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–586.

Andrews, D. W., M. Moreira, and J. Stock (2004): "Optimal Invariant Similar Tests for Instrumental Variables Regression," *NBER working paper 0299*.

Andrews, D. W. and J. H. Stock (2007): "Testing with many weak instruments," *Journal of Econometrics*, 138, 24–46.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K. and J. Stock (2005): "Inference with Weak Instruments," *NBER working paper t0313*.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

Andrews, I. and A. Mikusheva (2016a): "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

——— (2016b): "A Geometric Approach to Nonlinear Econometric Models," *Econometrica*, 84, 1249–1264.

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earning?" *Quaterly Journal of Economics*, 106 (4), 979–1014.

Antoine, B. and P. Lavergne (2022): "Identification-robust nonparametric inference in a linear IV model," *Journal of Econometrics*.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20, 105–134.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Chernozhukov, V. and C. Hansen (2008): "The reduced form: A simple approach to inference with weak instruments," *Economics Letters*, 100, 68–71.

Darolles, S., J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

Dieterle, S. G. and A. Snell (2016): "A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument," *Labour Economics*, 42, 76–86.

Dreier, I. and S. Kotz (2002): "A note on the characteristic function of the t-distribution," *Statistics & Probability Letters*, 57, 221–224.

Dufour, J.-M. (2003): "Identification, weak instruments, and statistical inference in econometrics," *Canadian Journal of Economics/Revue Canadienne d'Economique*, 36, 767–808.

Dufour, J.-M. and M. Taamouti (2007): "Further results on projection-based inference in IV regressions with weak, collinear or missing instruments," *Journal of Econometrics*, 139, 133–153.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10, 487–526.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Guggenberger, P. and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory*, 21.

Hahn, J. and J. Hausman (2003): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review*, 93, 118–125.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995): *Continuous univariate distributions 2*, vol. 2, Wiley.

Jun, S. J. and J. Pinkse (2012): "Testing Under Weak Identification with Conditional Moment Restrictions," *Econometric Theory*, 28, 1229–1282.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2005): "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica*, 73, 1103–1123.

——— (2007): "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

Li, Q. and J. S. Racine (2006): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Mikusheva, A. (2010): "Robust confidence sets in the presence of weak instruments," *Journal of Econometrics*, 157, 236–247.

Moreira, H. and M. J. Moreira (2019): "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," *Journal of Econometrics*, 398–433.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Moreira, M. J. and G. Ridder (2017): "Optimal Invariant Tests in an Instrumental Variables Regression With Heteroskedastic and Autocorrelated Errors," *arXiv:1705.00231*.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Seifert, B., T. Gasser, and A. Wolf (1993): "Nonparametric Estimation of Residual Variance Revisited," *Biometrika*, 80 (2), 373–383.

Staiger, D. and J. H. Stock' (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65 (3), 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Stock, J. H. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*.

Vaart, A. W. v. d. (2007): *Asymptotic statistics*, Cambridge University Press.

Vaart, A. W. v. d. and J. A. Wellner (2000): *Weak convergence and empirical processes: with applications to statistics*, Springer.

Yin, J., Z. Geng, R. Li, and H. Wang (2010): "Nonparametric Covariance Model," *Statistica Sinica*, 20, 469–479.

A Proof of propositions

A.1 Proof of Proposition 3.1

The implication from H_0 to H_0^1 and from H_0^1 to H_0^2 are obvious so I prove the reverse implications. From H_0^1 to H_0 notice that

$$\mathbb{E}(y_i - x_i\beta_0|z_i) = \Pi(z_i)'(\beta - \beta_0)$$

Thus if $l = 1$ and $\mathbb{P}(\Pi(z_i) = 0) = 0$ then $\mathbb{E}(y_i - x_i\beta_0|z_i)$ implies $\beta = \beta_0$.

From H_0^2 to H_0^1 first let $\mathbb{E}(y_i - x_i'\beta_0|z_i) = r(z_i) = max\{r(z_i); 0\} - max\{-r(z_i); 0\} = r_1(z_i) - r_2(z_i)$. Then

$$\mathbb{E}((y_i - x_i'\beta_0)e^{it'z_i}) = \mathbb{E}(r_1(z_i)e^{it'z_i}) - \mathbb{E}(r_2(z_i)e^{it'z_i}) = \int r_1(s)d\mathbb{P}_z(s) - \int r_2(s)d\mathbb{P}_z(s)$$

Next for $j = 1, 2$ define the probability measures $P_j(B) = \int_B r_j(s)d\mathbb{P}_z(s)/\mathbb{E}(r_j(z))\ \forall B \in \mathcal{B}(\mathbb{R}^k)$. This gives

$$\begin{aligned}
\forall t,\ \mathbb{E}((y_i - x_i'\beta_0)e^{it'z}) &= \mathbb{E}(r_1(z_i)) \int e^{it's}d\mathbb{P}_1(s) - \mathbb{E}(r_2(z_i)) \int e^{it's}d\mathbb{P}_2(s) = 0 \\
&\Rightarrow \mathbb{E}(r_1(z_i)) = \mathbb{E}(r_2(z_i)) \text{ by taking } t = 0 \\
&\Leftrightarrow \int e^{it's}d\mathbb{P}_1(s) - \int e^{it's}d\mathbb{P}_2(s) = 0\ \forall t \\
&\Leftrightarrow \forall B \in \mathcal{B}(\mathbb{R}^k),\ \mathbb{P}_1(B) - \mathbb{P}_2(B) = 0 \\
&\Leftrightarrow \forall B \in \mathcal{B}(\mathbb{R}^k),\ \int_B r(s)\mathbb{P}_z(s) = 0
\end{aligned}$$

Taking the event $B^+ = \{s : r(s) > 0\}$ and $B^- = \{s : r(s) < 0\}$ gives $\int_{B^+} r(s)d\mathbb{P}_z(s) = \mathbb{E}(r(z_i)1_{r(z_i)>0}) = 0 \Leftrightarrow \mathbb{P}(r(z_i) > 0) = 0$ and $\int_{B^-} r(s)d\mathbb{P}_z(s) = 0 \Leftrightarrow \mathbb{P}(r(z_i) < 0) = 0$, which imply that $\mathbb{P}(r(z_i) = 0) = 1 \Leftrightarrow r(z_i) = \mathbb{E}(y_i - x_i\beta_0|z_i) = 0$ a.s. hence H_0^2 implies H_0^1.

Equivalences between H_0^2, H_0^3 and H_0^4 are easily established using properties of positive functions and integrals.

A.2 Proof of Proposition 4.1

To prove that KICM is χ^2_l conditionally distributed first use the eigendecomposition of P_{WT}

$$P_{WT} = H \begin{pmatrix} I_l & 0_{l\times(n-l)} \\ 0_{(n-l)\times l} & 0_{(n-l)\times(n-l)} \end{pmatrix} H'$$

where $H = \begin{pmatrix} WT(T'W^2T)^{-1/2} & M_{WT}A(A'M_{WT}A)^{-1/2} \end{pmatrix}$, and $A = \begin{pmatrix} I_{n-l} \\ 0_{l\times(n-l)} \end{pmatrix}$, H is the orthogonal eigenvector matrix of P_{WT}. This allows to rewrite KICM as the sum of l components

$$\text{KICM} = \sum_{i=1}^{l} (S'H)_i^2$$

Second, I prove that $S'H \sim \mathcal{N}(0, I_n)$ conditionally on z. Under the null $H_0 : \beta = \beta_0$, $\mathbb{E}(S_i|z_i) = 0$ and $\text{Var}(S_i|z_i) = 1$ so that

$$\text{Cov}(S_i, T_i|z_i) = \mathbb{E}(S_iT_i|z_i) = \frac{b_0'\mathbb{E}(Y_iY_i')\Omega^{-1}A_0(A_0'\Omega^{-1}A_0)^{-1/2}}{\sqrt{b_0'\Omega b_0}} = 0_l$$

Thus conditionally on z_i $(S_i, T_i) \sim \mathcal{N}((0, \mathbb{E}(T_i)), I_{l+1})$. Consequently, under the null and conditionally on z_i $S_i \perp\!\!\!\perp T_i$ which implies that $S \perp\!\!\!\perp T$. In turn P is a function of z through W and of T therefore

$$\begin{aligned}
&\mathbb{E}(S'H|z,T) = \mathbb{E}(S|z)'H = 0_n \\
\Rightarrow &\text{Cov}((S'H)_i, (S'H)_j|z,T) = \mathbb{E}(e_i'HSS'He_j|z) = e_iH'\mathbb{E}(SS'|z)He_j = 1_{i=j} \\
\Rightarrow &\text{Var}(S'H|z,T) = H'\mathbb{E}(SS'|z)H = H'H = I_n \\
\Rightarrow &S'H \sim \mathcal{N}(0, I_n)|z,T \\
\Rightarrow &S'H \sim \mathcal{N}(0, I_n)
\end{aligned}$$

where e_j denotes a vector of size n equal to zero in all elements expect in coordinate j where it is equal to 1 and $1_{1=j}$ is an indicator function which equals one only when $i = j$. Consequently, KICM is equal to a sum of l independent squared standard normal so KICM follows a χ^2_l under $H_0 : \beta = \beta_0$.

B Proof of Theorem 4.3

Throughout the proof assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii), and D(iii) are maintained. The proof is divided in five parts, the first introduces notations and some

useful matrix results, the second presents the decomposition of the KICM statistic used in the proof, the third proves the convergence of the random processes which compose KICM, the fourth characterizes the limits of these random processes, and the fifth proves validity by contradiction.

B.1 Notations and matrix results

Denote by $o_{\mathbb{P}}(1)$ and $O_{\mathbb{P}}(1)$ the small o in probability and big O in probability notations for degenerate in probability and bounded in probability: If $X = o_{\mathbb{P}}(1)$ then $\forall \varepsilon > 0\ \mathbb{P}(|X| > \varepsilon) \to 0$. If $X = O_{\mathbb{P}}(1)$ then $\forall \varepsilon > 0\ \exists M > 0 : \mathbb{P}(|X| > M) \leqslant \varepsilon$. Denote by $o_{\underline{\mathbb{P}}}(1)$ and $O_{\underline{\mathbb{P}}}(1)$ the uniform counterparts of $o_{\mathbb{P}}(1)$ and $O_{\mathbb{P}}(1)$: If $X = O_{\underline{\mathbb{P}}}(1)$ then X is uniformly bounded in probability and $\forall \epsilon > 0\ \exists M : \underset{\mathbb{P}\in\mathcal{P}}{sup}\ \mathbb{P}(|X| > M) \leqslant \epsilon$. If $X = o_{\underline{\mathbb{P}}}(1)$ then X is uniformly degenerate in probability and $\forall \varepsilon > 0,\ \underset{\mathbb{P}\in\mathcal{P}}{sup}\ \mathbb{P}(|X| > \varepsilon) \to 0$. For any random object X then denote by $X \neq 0$ the condition $\mathbb{P}(X = 0) < 1$ or equivalently $\mathbb{P}(X \neq 0) > 0$. In addition, for some square matrix X denote by $\lambda_{min}(X)$ and $\lambda_{max}(X)$ its smallest and biggest eigenvalue respectively.

Before starting the proof here are four useful matrix results which are used many times over: For any invertible matrices X and Y, for any full rank matrix A

$$X^{-1} - Y^{-1} = X^{-1}(Y - X)Y^{-1} \tag{B.6}$$

$$X - Y = \frac{1}{2}(X^{1/2} + Y^{1/2})(X^{1/2} - Y^{1/2}) + \frac{1}{2}(X^{1/2} - Y^{1/2})(X^{1/2} + Y^{1/2}) \tag{B.7}$$

$$(A'XA)^{-1} \leqslant (A'A)^{-1}\ \lambda_{min}(X)^{-1} \tag{B.8}$$

$$((A'XA)^{-1/2} + (A'YA)^{-1/2})^{-1} \leqslant (A'A)^{1/2}(\lambda_{max}(X)^{-1/2} + \lambda_{max}(Y)^{-1/2})^{-1}\} \tag{B.9}$$

$$((A'X^{-1}A)^{-1/2} + (A'Y^{-1}A)^{-1/2}) \geqslant (A'A)^{-1/2}(\lambda_{min}(X)^{1/2} + \lambda_{min}(Y)^{1/2})\} \tag{B.10}$$

B.2 Preliminary decomposition

As defined in section 3.3

$$T_{ih} = (A_0'\Omega_i^{-1}A_0)^{-1/2}A_0'\Omega_i^{-1}Y_i$$

Thus each component j of T_{ih} is a weighted sum of the components of Y_i. Thus without loss of generality

$$T_{ih,j} = \sum_{l'=1}^{l} w_{ij,l'}x_{il'} + w_{ij,y}y_i = \sum_{l'=1}^{l}(w_{ij,l'} + w_{ij,y}\beta_{0l'})x_{il'} + w_{ij,y}u_i \equiv \sum_{l'=1}^{l} w^*_{ij,l'}x_{il'} + w_{ij,y}u_i$$

where $T_{ih,j}$ is the j-th coordinate of T_{ih}, $w_{ij,l'} \in \mathbb{R}$ is the weight associated to $x_{il'}$ through x_i, $w_{ij,y} \in \mathbb{R}$ the weight associated to y_i, and $(w^*_{ij,l'} \in \mathbb{R}$ is the true weight associated to $x_{il'}$. These

weights are functions of z_i and of the null β_0 and note that $\forall j\ T_{ih,j} \neq 0$ because $\text{Var}(T_i) = I_n$, thus $\forall j\ (w^*_{j1i}, \ldots, w^*_{jli}, w_{ij,y}) \neq 0_{l+1}$. Therefore define $(a^*_j)_{j=1}^l$ where

$$a^*_j = \begin{cases} a & \text{if} \quad (w^*_{ij,l'})_{l'=1}^l \neq 0_l \\ 1 & \text{o.w} \end{cases}$$

a^*_j does not represent identification strength of β_j but instead characterize the limit (which exists) of $\frac{1}{\sqrt{n}} \sum_{i=1}^n T_{ih,j}$. Additionally a^*_j does not only depend on instruments' strength a, but also on the conditional covariance between x_{ij} and $(x_{ij'})_{j'=1}^l$, and on β_0. To get an heuristic idea of why a^*_j is introduced, notice that because $\Pi(\cdot) = n^{-a}C(\cdot)$ if $a^*_j > 1/2$ then $\frac{1}{\sqrt{n}} \sum_{i=1}^n T_{ih,j}$ will converge towards a centered Normal distribution, if $a^*_j = 1/2$ then $\frac{1}{\sqrt{n}} \sum_{i=1}^n T_{ih,j}$ will converge towards a non-centered distribution, and if $a^*_j < 1/2$ then $\frac{n^{a^*_j - 1/2}}{\sqrt{n}} \sum_{i=1}^n T_{ih,j}$ will converge in probability towards a certain expectation. Hence KICM rewrites as follows

$$\text{KICM}_h = S_h' W T_h (T_h' W^2 T_h)^{-1} T_h' W S_h = S_h' W T_h N_C^* (N_C^* T_h' W^2 T_h N_C^*)^{-1} N_C^* T_h' W S_h$$

where $N_C^* = \begin{pmatrix} n^{a^*_1 - 1/2} 1_{a^*_1 < 1/2} + 1_{a^*_1 \geqslant 1/2} & 0 & \cdots & 0 \\ 0 & n^{a^*_2 - 1/2} 1_{a^*_2 < 1/2} + 1_{a^*_2 \geqslant 1/2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n^{a^*_l - 1/2} 1_{a^*_l < 1/2} + 1_{a^*_l \geqslant 1/2} \end{pmatrix}$

N_C^* is a $l \times l$ diagonal matrix, it is a theoretical tool which will allow to properly characterize the limit of KICM, given any instruments' strength, any covariance structure $\Omega(\cdot)$ and any null β_0.

I can then rewrite the components of KICM_h. $S_h' W T_h N_C^*$ is the integral of the product of the sums of $(S_{ih})_{i=1}^n$ and of $(T_{ih})_{i=1}^n$

$$\begin{aligned} S_h' W T_h N_C^* &= \frac{1}{n} \sum_{i,j} S_{ih} T_{jh}' N_C^* w(z_i - z_j) = \int \frac{1}{n} \sum_{i,j} S_{ih} T_{jh}' N_C^* e^{it'(z_i - z_j)} \mu(t) \\ &= \int \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n S_{ih} e^{it' z_i} \right) \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n T_{jh}' N_C^* e^{-it' z_j} \right) \mu(t) \end{aligned}$$

$N_C^* T_h' W^2 T_h N_C^*$ also rewrites. I reformulate the elements of the matrix W^2 first

$$\begin{aligned}(W^2)_{ij} &= \frac{1}{n^2}\sum_{m=1}^{n}\int_{\mathbb{R}^k} e^{it'(z_i - z_m)}d\mu(t)\int_{\mathbb{R}^k} e^{is'(z_m - z_j)}d\mu(s) \\ &= \frac{1}{n^2}\int\int e^{it'z_i}e^{-is'z_j}\left(\sum_{m=1}^{n} e^{i(s-t)z_m}\right)d\mu(t)d\mu(s)\end{aligned}$$

$$\Rightarrow N_C^* T_h' W^2 T_h N_C^* = \sum_{i,j} N_C^* T_{ih} T_{jh}' N_C^* (W^2)_{ij}$$

$$N_C^* T_h' W^2 T_h N_C^* = \int\int \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} N_C^* T_{ih} e^{it'z_i}\right)\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} T_{jh}' N_C^* e^{-is'z_j}\right)\left(\frac{1}{n}\sum_{m=1}^{n} e^{i(s-t)z_m}\right)d\mu(t)d\mu(s)$$

Thereforedefine the processes which enter the integrals and characterize KICM

$$\begin{aligned}G_S(t,\Omega) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(b_0'\Omega(z_i)b_0)^{-1/2}b_0'Y_i e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ih}e^{it'z_i} \\ G_T(t,\Omega) &= \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n}(A_0'\Omega^{-1}(z_i)A_0)^{-1/2}A_0'\Omega^{-1}(z_i)Y_i e^{-it'z_i} = \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n} T_{ih}e^{-it'z_i} \\ G_z(t) &= \frac{1}{n}\sum_{i=1}^{n} e^{it'z_i}\end{aligned}$$

So KICM$_h$ rewrites as

$$\int_{\mathbb{R}^k} G_S(t,\Omega)G_T'(t,\Omega)d\mu(t)\left(\int_{\mathbb{R}^k} G_T(-t,\Omega)G_T'(s,\Omega)G_z(s-t)d\mu(t)\right)^{-1}\int_{\mathbb{R}^k} G_S(t,\Omega)G_T(t,\Omega)d\mu(t)$$

Lastly, define the conditionally normal version of G_S and G_T

$$\begin{aligned}G_S^*(t,\Omega) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(b_0'\Omega(z_i)b_0)^{-1/2}b_0'Y_i^* e^{it'z_i}, \quad Y_i^*|z_i \overset{iid}{\sim} \mathcal{N}(\mathbb{E}(Y_i|z_i),\Omega(z_i)) \\ G_T^*(t,\Omega) &= \frac{N_C^*}{\sqrt{n}}\sum_{i=1}^{n}(A_0'\Omega^{-1}(z_i)A_0)^{-1/2}A_0'\Omega^{-1}(z_i)Y_i^* e^{-it'z_i}\end{aligned}$$

B.3 Donsker property and asymptotic equicontinuity

From the above decomposition in order for KICM to converge, the processes G_S, G_T, G_S^*, G_T^* must uniformly converge to Gaussian processes, ie to be Donsker, and G_z must uniformly converge to a function, ie to be Glivenko-Cantelli. In addition, in order to make the difference

between the feasible KICM test statistic and the KICM test statistic the processes G_S, G_S^*, G_T and G_T^* also need to be asymptotically equicontinuous in Ω.

To obtain these results the following families of functions

$$\mathcal{F}_S = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (b_0'\Omega b_0)^{-1/2} b_0' d \; e^{it'c}, t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_T = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (A_0'\Omega^{-1}A_0)^{-1/2} A_0'\Omega^{-1} d \; e^{it'c}, t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_z = \{c \in \mathbb{R}^k \mapsto e^{it'c}, t \in \mathbb{R}^k\}$$

need an envelope F which satisfies $\mathbb{E}(F^2(y_i,x_i,z_i)1_{F(y_i,x_i,z_i)\geqslant M}) \underset{M\to+\infty}{\to} 0$ and a bounded uniform entropy integral (BUEI). If that's the case then by Theorem 2.8.3 in Vaart and Wellner (2000) $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ are uniformly Donsker and pre-Gaussian and thus by Theorem 2.8.2 from Vaart and Wellner (2000) processes G_S, G_T, G_S^* and G_T^* weakly converge to Gaussian processes and are uniformly asymptotically equicontinuous, and G_z weakly converge to a function. Asymptotic equicontinuity implies that for any $\Omega \in \mathcal{O}$ with consistent estimator $\hat{\Omega}$ as per assumption D the difference between the feasible $G_S(\cdot,\Omega)$ and $G_T(\cdot,\hat{\Omega})$ vanishes uniformly in probability

$$\forall \varepsilon > 0, \; \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta=\beta_0} \mathbb{P}(|G_S(\cdot,\Omega) - G_S(\cdot,\hat{\Omega})| > \varepsilon) \to 0$$

$$\forall j = 1, \cdots, l, \; \forall \varepsilon > 0, \; \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta=\beta_0} \mathbb{P}(|G_{Tj}(\cdot,\Omega) - G_{Tj}(\cdot,\hat{\Omega})| > \varepsilon) \to 0$$

Here assumption D(iii) and D(iv) was used, specifically the fact that $\hat{\Omega}$ converge uniformly towards Ω in the L_2 sense and that $\hat{\Omega}$ is uniformly in $\mathcal{O}$ at the limit.

To show that $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ have a "bounded" envelope and are BUEI, I first consider

$$\mathcal{F}_{cos,S} = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (b_0'\Omega b_0)^{-1/2} b_0' d \; cos(t'c), t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$
$$\mathcal{F}_{cos,T} = \{(c,d) \in \mathbb{R}^k \times \mathbb{R}^{l+1} \mapsto (A_0'\Omega^{-1}A_0)^{-1/2} A_0'\Omega^{-1} d \; cos(t'c), t \in \mathbb{R}^k, \Omega \in \mathcal{O}\}$$

Note that $|cos(.)| \leqslant 1$, and that from assumption D(ii) $\exists(\underline{\lambda},\bar{\lambda}) : \underline{\lambda} I_{l+1} \leqslant \Omega(\cdot) \leqslant \bar{\lambda} I_{l+1}$ where $\underline{\lambda} > 0$ and $\bar{\lambda} < +\infty$. Thus for any $(t,\Omega) \in \mathbb{R}^k \times \mathcal{O}$, for any $A \in \mathbb{R}^{(l+1)\times l}$ full rank, for any $b \neq 0_{l+1}$, for any $K \in \mathbb{R}^{l+1}$

$$|(b_0'\Omega b_0)^{-1/2} b_0' d \; cos(t'c)| \leqslant \underline{\lambda}^{-1/2} ||b_0||_2^{-1/2} |b_0' d|$$
$$|(A_0'\Omega^{-1}A_0)^{-1/2} A_0'\Omega^{-1} d \; cos(t'c)|_1 \leqslant \bar{\lambda}^{1/2}\underline{\lambda}^{-1} |(A_0'A_0)^{-1/2} A_0 d|_1$$

Thus an envelope for $\mathcal{F}_{cos,S}$ is $F_{\mathcal{F}_{cos,S}} : d \mapsto \underline{\lambda}^{-1/2} ||b_0||_2^{-1/2} |b_0' d|$, $F_{\mathcal{F}_{cos,S}}(Y_i)$ is square integrable by A(iii), in addition $1_{F_{\mathcal{F}_{cos,S}}(Y_i)>M} \underset{M\to+\infty}{\to} 0$, thus by the dominated convergence theorem (DCT) $\mathbb{E}(F_{\mathcal{F}_{cos,S}}(Y_i)^2 1_{F\geqslant M}) \underset{M\to+\infty}{\to} 0$. Envelopes by coordinate for $\mathcal{F}_{cos,T}$ are $F_{l',\mathcal{F}_{cos,T}} : d \mapsto$

$\bar{\lambda}^{1/2}\underline{\lambda}^{-1}|e_j'(A_0'A_0)^{-1/2}A_0d|$ with $l' = 1,\dots,l$, using previous arguments by the DCT they also satisfy the condition. Using similar arguments $\mathcal{F}_S$, $\mathcal{F}_T$ and $\mathcal{F}_z$ also satisfy this condition and have a "bounded" envelope.

Next, by assumption D(ii) $\mathcal{O}$ has finite covering number so it is BUEI. The fact that the composite of Ω are BUEI still needs to be proven. To do that, I prove that the functions $\Omega \mapsto (A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}d$ and $\Omega \mapsto (b_0'\Omega b)^{-1/2}b_0'd$ are Lipschitz and bounded in Ω $\forall d \in \mathbb{R}^{l+1}$ so that by Lemma 9.14 in Kosorok (2008) the families $\{d \mapsto (b_0'\Omega b_0)^{-1/2}b_0'd, \Omega \in \mathcal{O}\}$ and $\{d \mapsto (A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}d, \Omega \in \mathcal{O}\}$ are BUEI.

Boundedness is trivial from the fact that any covariance function in $\mathcal{O}$ has eigenvalues in a bounded subset of $\mathbb{R}_*^+$. First, I deal with Lipschitz continuity of the first function

$$\begin{aligned}
\frac{|b_0'd(b_0'\Omega_1b_0)^{-1/2} - b_0'd(b_0'\Omega_2b_0)^{-1/2}|}{|(b_0'\Omega_1b_0)^{-1} - (b_0'\Omega_2b_0)^{-1}|} &= \frac{|b_0'd(b_0'\Omega_1b_0)^{-1/2} - b_0'd(b_0'\Omega_2b_0)^{-1/2}|}{|(b_0'\Omega_1b_0)^{-1/2} - (b_0'\Omega_2b_0)^{-1/2}|\ |(b_0'\Omega_1b_0)^{-1/2} + (b_0'\Omega_2b_0)^{-1/2}|} \\
&= |b_0'd|\ |(b_0'\Omega_1b_0)^{-1/2} + (b_0'\Omega_2b_0)^{-1/2}|^{-1} \\
&\leqslant |b_0'd|\ ||b||_2(\lambda_{max}(\Omega_1)^{-1/2} + \lambda_{max}(\Omega_2)^{-1/2})^{-1} \\
&\leqslant |b_0'd|\ ||b||_2\ \frac{\sqrt{\bar{\lambda}}}{2} \equiv \tilde{K}
\end{aligned}$$

where the 3rd line is obtained by (B.9) and the last line by assumption D(ii), and due to the fact that eigenvalues of any $\Omega \in \mathcal{O}$ are bounded in $\mathbb{R}_*^+$. This implies Lipchitzness of $\Omega \mapsto (b_0'\Omega b_0)^{-1/2}b_0'd$ using (B.8)

$$\begin{aligned}
|b_0'd(b_0'\Omega_1b_0)^{-1/2} - b_0'd(b_0'\Omega_2b_0)^{-1/2}| &\leqslant \tilde{K}\ |(b_0'\Omega_1b_0)^{-1} - (b_0'\Omega_2b_0)^{-1}| \\
&\leqslant \tilde{K}\ ||(b_0'\Omega_2b_0)^{-1}(b_0'\Omega_2b_0 - b_0'\Omega_1b_0)(b_0'\Omega_1b_0)^{-1}||_2 \\
&\leqslant \tilde{K}\ ||b||_2^{-2}\lambda_{min}(\Omega_1)^{-1}\lambda_{min}(\Omega_2)^{-1}||b_0'(\Omega_2 - \Omega_1)b||_2 \\
&\leqslant \tilde{K}\ ||b||_2^{-2}\underline{\lambda}^{-2}||b_0'(\Omega_2 - \Omega_1)b||_2 \\
&\leqslant \frac{\tilde{K}}{||b||_2\ \underline{\lambda}^2}||\Omega_1 - \Omega_2||_2 \\
&= \frac{|b_0'd|\sqrt{\bar{\lambda}}}{2\underline{\lambda}^2}||\Omega_1 - \Omega_2||_2
\end{aligned}$$

Next without loss of generality assume that $\Omega_1 > \Omega_2$. This implies that

$$\Omega_1^{-1} < \Omega_2^{-1} \Rightarrow\ A_0'\Omega_1^{-1}A_0 < A_0'\Omega_2^{-1}A_0 \Rightarrow (A_0'\Omega_1^{-1}A_0)^{-1/2} > (A_0'\Omega_2^{-1}A_0)^{-1/2}$$

Then using (B.10) I obtain that

$$\frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2}) \geqslant \frac{1}{2}(A_0'A_0)^{-1/2}(\lambda_{min}(\Omega_1)^{-1/2} + \lambda_{min}(\Omega_2)^{-1/2})^{-1} \geqslant (A_0'A_0)^{-1/2}\sqrt{\underline{\lambda}}$$

It then follows using (B.7)

$$
\begin{aligned}
||(A_0'\Omega_1^{-1}A_0)^{-1} - (A_0'\Omega_2^{-1}A_0)^{-1}||_2 &= ||\frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2})((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}) \\
&+ \frac{1}{2}((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2})((A_0'\Omega_1^{-1}A_0)^{-1/2} + (A_0'\Omega_2^{-1}A_0)^{-1/2})||_2 \\
&\geqslant \sqrt{\underline{\lambda}}\, ||(A_0'A_0)^{-1/2}((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}) \\
&+ ((A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2})(A_0'A_0)^{-1/2}||_2 \\
&= 2\sqrt{\frac{\underline{\lambda}}{\lambda_{max}(A_0'A_0)}}||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2 \\
&\equiv \tilde{\tilde{K}}||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2
\end{aligned}
$$

Using (B.6) the above inequality implies the following

$$
\begin{aligned}
||(A_0'\Omega_1^{-1}A_0)^{-1/2} - (A_0'\Omega_2^{-1}A_0)^{-1/2}||_2 &\leqslant \tilde{\tilde{K}}^{-1}\, ||(A_0'\Omega_1^{-1}A_0)^{-1} - (A_0'\Omega_2^{-1}A_0)^{-1}||_2 \\
&\leqslant \tilde{\tilde{K}}^{-1}\, ||(A_0'\Omega_1^{-1}A_0)^{-1}||_2\, ||(A_0'\Omega_2^{-1}A_0)^{-1}||_2\, ||A_0||_2^2\, ||\Omega_1 - \Omega_2||_2 \\
&\leqslant \tilde{\tilde{K}}^{-1}\, ||(A_0'A_0)^{-1}||_2^2\, \bar{\lambda}^2\, ||\Omega_1 - \Omega_2||_2
\end{aligned}
$$

Then using the triangular inequality and previous arguments it follows that $\Omega \mapsto (A_0'\Omega^{-1}A_0)^{-1/2}A_0'\Omega^{-1}$ is Lipschitz $\forall d \in \mathbb{R}l + 1$ and for some strictly positive constant $\tilde{\tilde{\tilde{K}}}$

$$
\begin{aligned}
||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_1^{-1}K - (A_0'\Omega_2^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 &\leqslant ||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_1^{-1}K - (A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 \\
&+ ||(A_0'\Omega_1^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K - (A_0'\Omega_2^{-1}A_0)^{-1/2}A_0'\Omega_2^{-1}K||_2 \\
&\leqslant |\tilde{\tilde{\tilde{K}}}|\, ||\Omega_1 - \Omega_2||_2
\end{aligned}
$$

As for the class $\{c \mapsto t'c, t \in \mathbb{R}^k\}$ it has Vapnick-Cervonenkis index $k+1$ hence by Sauer's Lemma (Vaart (2007)) it is BUEI. From there as the cosinus function is also bounded and Lipschitz continuous then the class $\{c \mapsto cos(t'c), t \in \mathbb{R}^k\}$ is BUEI by 9.14 in Kosorok (2008). The Lipschitz-continuity in t can be proven with the mean value theorem in the following way

$$
\begin{aligned}
\frac{|cos(t_1'c) - cos(t_2'c)|}{||t_1 - t_2||_2} &\leqslant \frac{|cos(t_1'c) - cos(t_2'c)|}{|(t_1 - t_2)'c|}\, \frac{|(t_1 - t_2)'c|}{||t_1 - t_2||_2} \\
&\leqslant ||c||_2 \frac{|cos(t_1'c) - cos(t_2'c)|}{|(t_1 - t_2)'c|} \leqslant ||c||_2
\end{aligned}
$$

Then by Theorem 9.15 in Kosorok (2008) the "product" of 2 BUEI families is BUEI so that $\mathcal{F}_{cos,T}$ and $\mathcal{F}_{cos,S}$ are BUEI. Replacing the cosinus function by the sinus function keep these two families BUEI therefore by applying Lemma 9.14 from Kosorok (2008), the families $\mathcal{F}_T$, $\mathcal{F}_S$ and $\mathcal{F}_z$ are BUEI.

B.4 Limits of G_S, G_T and G_z

Next, from the previous results G_S and G_S^* converge uniformly towards Gaussian processes if demeaned whereas G_z converges uniformly towards a function with values in $\mathbb{C}$. As for G_T and G_T^*, depending on instruments' strength a it can be decomposed as the sum of different terms, some of which are uniformly degenerate, some of which converge uniformly towards functions with values in $\mathbb{C}$, and some of which converge uniformly towards Gaussian processes. Note that the families of functions over which G_S, G_T, G_ϵ, G_ζ, and G_z are defined such as $\mathcal{F}_T$ and $\mathcal{F}_S$ are Donsker thus also Glivenko-Cantelli. These limits are characterized in the following way:

1. Regarding G_S recall $G_S(t,\Omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ih}e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i'b_0(b_0'\Omega_i b_0)^{-1/2}e^{it'z_i}$, therefore under the null the and $\forall t \in \mathbb{R}^k$, $\forall \Omega \in \mathcal{O}$ the asymptotic mean of G_S writes

$$\mathbb{E}(S_{ih}e^{it'z_i}) = \mathbb{E}(\mathbb{E}(y_i - x_i'\beta_0|z_i)(b_0'\Omega_i b_0)^{-1/2}e^{it'z_i}) = \mathbb{E}(\epsilon_i e^{it'z_i}) = 0$$

   And $\forall (t,\tilde{t})$, $\forall \Omega \in \mathcal{O}$, the asymptotic covariance of G_S writes

$$\mathrm{Var}(S_{ih}e^{it'z_i}) = \mathbb{E}(\mathbb{E}((y_i - x_i'\beta_0)^2|z_i)(b_0'\Omega_i b_0)^{-1}e^{2it'z_i}) = \mathrm{Var}(u_i e^{it'z_i}) = \mathbb{E}(e^{2it'z_i})$$
$$\mathrm{Cov}(S_{ih}e^{it'z_i}, S_{ih}e^{i\tilde{t}'z_i}) = \mathbb{E}(e^{i(t+\tilde{t})'z_i})$$

   Consequently, $\forall \Omega \in \mathcal{O}$ the limit of $G_S(\cdot,\Omega)$ is a Gaussian process with, under the null, a constant mean equal to 0 and a covariance function $(t,\tilde{t}) \mapsto \mathbb{E}(e^{i(t+\tilde{t})'z_i})$. Note that under any alternative $H_1 : \beta \neq \beta_0$ the asymptotic equicontinuity result still holds, ie the difference between $G_S(\cdot,\Omega)$ and $G_S(\cdot,\hat{\Omega})$ vanishes, however G_S does not converge towards the aforementioned Gaussian process. Characterization of the limit of G_S in Ω is unecessary for the proof, G_S is considered a function of Ω only in order to prove convergence of the feasible process $G_S(\cdot,\hat{\Omega})$.

   Regarding G_S^*, it has the same limit as G_S because the conditional normality doesn't play a role asymptotically, indeed $\mathbb{E}(Y_i) = \mathbb{E}(Y_i^*)$ and $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_i^*)$. Thus for any $\Omega \in \mathcal{O}$ both $G_S(\cdot,\Omega)$ and $G_S^*(\cdot,\Omega)$ converge uniformly towards the same complex Gaussian process. This implies that the distance between the 2 shrinks asymptotically uniformly by definition of weak convergence.

2. Regarding G_z its limit is the characteristic function of z_i $t \mapsto \mathbb{E}(e^{it'z_i})$.

3. Regarding G_T, first recall

$$T_{ih,j} = \sum_{l'=1}^{l} w^*_{ij,l'}x_{il'} + w_{ij,y}u_i = \sum_{l'=1}^{l} w^*_{ij,l'}(\frac{C(z_i)_{l'}}{n^a} + v_{il'}) + w_{ij,y}u_i = n^{-a}w_{ij}^{*\prime}C(z_i) + w_{ij}^{*\prime}v_i + w_{ij,y}u_i$$

Therefore for some $j = 1, \dots, l$ G_{Tj} rewrites

$$G_{Tj}(t, \Omega) = \frac{n^{a_j^* - 1/2} 1_{a_j^* < 1/2} + 1_{a_j^* \geqslant 1/2}}{\sqrt{n}} \sum_{i=1}^{n} T_{ih,j} e^{-it' z_i}$$

where $a_j^* = \begin{cases} a & \text{if } (w_{ij,l'}^*)_{l'=1}^{l} \neq 0_l \\ 1 & \text{o.w} \end{cases}$ so if $a_j^* \geqslant 1/2$ then G_{T_j} writes

$$\begin{aligned} G_{T_j}(t, \Omega) = &\sum_{l'=1}^{l} \frac{1}{n^{1/2+a}} \sum_{i=1}^{n} w_{ij,l'}^* C(z_i)_{l'} e^{-it' z_i} + \sum_{l'=1}^{l} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,l'}^* v_{il'} e^{-it' z_i} \\ &+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,y} u_i e^{-it' z_i} \end{aligned}$$

And if $a_j^* < 1/2$ it must be that $a_j^* = a$ so G_{T_j} writes

$$\begin{aligned} G_{T_j}(t, \Omega) = &\sum_{l'=1}^{l} \frac{1}{n} \sum_{i=1}^{n} w_{ij,l'}^* C(z_i)_{l'} e^{-it' z_i} + \sum_{l'=1}^{l} \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,l'}^* v_{il'} e^{-it' z_i} \\ &+ \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,y} u_i e^{-it' z_i} \end{aligned}$$

It follows then that if $a_j^* > 1/2$ then, either $\forall l'$ $w_{ij,l'} = 0$ and $a_j^* = 1$, either $\exists l'$: $\mathbb{E}(w_{ij,l'} C(z_i)_{l'}) \neq 0$ and $a_j^* = a > 1/2$. Thus if $a_j^* > 1/2$

$$\sum_{l'=1}^{l} \frac{1}{n^{1/2+a}} \sum_{i=1}^{n} w_{ij,l'}^* C(z_i)_{l'} e^{-it' z_i} = o_{\underline{\mathbb{P}}}(1)$$

This implies that $G_{T_j}(\cdot, \Omega)$ will converge to a Gaussian process with constant mean 0 and covariance function $(t, s) \mapsto \mathbb{E}(e^{-i(t+s)' z_i})$ because $\sum_{l'=1}^{l} \mathbb{E}(w_{ij,l'} v_{il'} e^{-t' z_i}) = \mathbb{E}(w_{ij,y} u_i e^{-it' z_i}) = 0$ by the law of iterated expectations and because T_{ih} has variance identity conditionally on z_i.

If $a_j^* = 1/2$ then $G_{T_j}(\cdot, \Omega)$ will converge towards a non-centered Gaussian process with mean $t \mapsto \sum_{l=1}^{l} {}' \mathbb{E}(w_{ij,l'}^* C(z_i) e^{-itz_i})$ and covariance function $(t, s) \mapsto \mathbb{E}(e^{-i(t+s)' z_i})$.

If $a_j^* < 1/2$ then $\exists l' : w_{ij,l'}^* \neq 0$ and $a_j^* = a$. Then

$$\sum_{l'=1}^{l} \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{ij,l'}^* v_{il'} e^{-it' z_i} + \frac{n^{a-1/2}}{\sqrt{n}} \sum_{i=1}^{n} w_{jyi}^* u_i e^{-it' z_i} = o_{\underline{\mathbb{P}}}(1)$$

Therefore $G_{T_j}(\cdot, \Omega)$ will converge towards $t \mapsto \sum_{l'=1}^{l} \mathbb{E}(w^*_{ij,l'} C(z_i)_{l'} e^{-it' z_i})$ for any $\Omega(\cdot) \in \mathcal{O}$.

Regarding the whole vector T_i, because T_i has conditional variance identity (even under the alternative), $G_T(\cdot, \Omega) = (G_{T_1} \ \cdots \ G_{T_l})$ will converge to a vector of random processes whose elements are uncorrelated and characterized by the limits G_{T_j} for $j = 1, \ldots, l$ which are characterized by the vector $(a^*_j)_{j=1}^{l}$.

Now with regards to G^*_T, it has the same limit as G_T for the same reason that G^*_S and G_S have the same limit: the conditional normality does not play a role anymore asymptotically, asymptotically what matters are the means and the covariances of the limitting processes which in the case G_T and G^*_T are the same. In conclusion, asymptotically the distance between $G_T(\cdot, \Omega)$ and $G^*_T(\cdot, \Omega)$ vanishes for any $\Omega \in \mathcal{O}$ uniformly.

B.5 Lipschitzness of KICM

Define f and g for any triple of functions (A, B, C) with inputs in $\mathbb{R}^k$ and values in a bounded subset of $\mathbb{C}$, $\mathbb{C}^l \setminus 0_l$ and $\mathbb{C}^+_*$ respectively

$$f(A, B) = \int_{\mathbb{R}^k} A(t)B(t)\mu(t), \quad g(B, C) = \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} B(t)B'(t)C(s-t)d\mu(s)d\mu(t)$$

Notice that $\text{KICM}_h = ||f(G_S, G_T)g(G_T, G_z)||_2^2$ so I shall prove that KICM_h is Lipschitz in G_S, G_T and G_z through f and g. In the following K is an unspecified constant.

- To prove Lipschitzness in A note that because (A, B, C) take values in bounded subset I can always find a supremum to $A(t)$ and $B(t)$ and an infimum to $\lambda_{min}(g(B, C))$

$$\begin{aligned} | \, ||f(A_1, B)g(B, C)^{-1/2}||_2^2 - ||f(A_2, B)g(B, C)^{-1/2}||_2^2 \, | &= | \int (A_1 - A_2)(t)B(t)' d\mu(t) g(B, C)^{-1} \\ &\quad \int (A_1 + A_2)(t)B(t)d\mu(t) \, | \\ &\leqslant ||A_1 - A_2||_\infty \int ||B(t)||_2 d\mu(t) \lambda_{min}(g(B, C))^{-1} \\ &\quad \int |A_1(t) + A_2(t)| \, ||B(t)||_2 d\mu(t) \\ &\leqslant K \times \sup_t |A_1(t) - B_2(t)| \end{aligned}$$

- To prove Lipschitzness in C I use result ((B.6))

$$\begin{aligned}||g(B,C_1)-g(B,C_2)||_2 &= \leqslant \lambda_{min}(g(B,C_1))^{-1}\lambda_{min}(g(B,C_2))^{-1}\\ &\times \int\int ||B(t)B(-s)'||_2 d\mu(t)d\mu(s)||C_1-C_2||_\infty\\ &\leqslant K\,||C_1-C_2||_\infty\end{aligned}$$

  Then Lipschitzness of $C \mapsto ||f(A,B)g(B,C)^{-1/2}||_2^2$ is established because it is a function of C only through g and because I can find an upper bound on $f(A,B)$.

- To prove Lipchitzness in B I express $|\ ||f(A,B_1)g(B_1,C)^{-1/2}||_2^2 - ||f(A,B_2)g(B_2,C)^{-1/2}||_2^2\ |$ as a sum of 3 components which depend on $f(A,B_1)-f(A,B_2)$ and $g(B_1,C)-g(B_2,C)$ then by the triangular inequality and for some positive constants (K_1,K_2)

$$\begin{aligned}&|\ ||f(A,B_1)g(B_1,C)^{-1/2}||_2^2 - ||f(A,B_2)g(B_2,C)^{-1/2}||_2^2\ |\\ &= |f(A,B_1)'(g(B_1,C)^{-1}-g(B_2,C)^{-1})f(A,B_1)\\ &\quad + f(A,B_2)'\left(g(B_1,C)^{-1}(f(A,B_1)-f(A,B_2)) - (g(B_2,C)^{-1}-g(B_1,C)^{-1})f(A,B_2)\right)|\\ &\leqslant K_1\,||g(B_1,C)^{-1}-g(B_2,C)^{-1}||_2 + K_2\,||f(A,B_1)-f(A,B_2)||_2\end{aligned}$$

  Reusing (B.6) and aforementioned arguments for the Lipchitzness in A and C of $(A,B,C) \mapsto\mapsto ||f(A,B)g(B,C)^{-1/2}||_2^2$ it is also Lipchitz in B.

The Lipschitzness of KICM_h in G_S and G_T allows for the difference between feasible KICM (KICM_{hf}) and unfeasible KICM (KICM_h) to vanish uniformly asymptotically.

B.6 KICM validity by contradiction

Let $q_{1-\alpha}$ be the $1-\alpha$ quantile of a χ^2_l then suppose that the theorem does not hold. The theorem not holding is equivalent to $\exists P$ such that $\beta = \beta_0$ and $\mathbb{P}(\text{KICM}_h > q_{1-\alpha}) > \alpha$ which implies that $\exists \delta > 0 : \mathbb{P}(\text{KICM}_h > q_{1-\alpha}) \geqslant \alpha + 2\delta$.

Next, I can find bounded subsets $C_S \subset \mathbb{C}_*$, $C_T \subset \mathbb{C}_*^l$ and $C_z \subset \mathbb{C}_*$ such that $\delta > \mathbb{P}(G_S \notin C_S, G_T \notin C_T, G_z \notin C_z) = 1 - \mathbb{P}(G_S \in C_S, G_T \in C_T, G_z \in C_z)$. Indeed, these C_S, C_T and C_z exist because G_S, G_T and G_z are uniformly bounded in probability and non-degenerate as shown in B.4. Note that, because the integral of a bounded random process with respect to a finite measure is bounded and the product of bounded random variables is bounded, the event $G_S \in C_S \cap G_T \in C_T \cap G_z \in C_z$ implies some event $\text{KICM}_h \leqslant C$ which is consistent with the fact that $\text{KICM}_h = O_{\mathbb{P}}(1) \neq o_{\underline{\mathbb{P}}}(1)$.

Next, let K_h be KICM_h but with Y_i assumed normal conditionally on z_i. Then as I saw before in 4.1 $\text{K}_h \sim \chi^2_l | z, T \Rightarrow \text{K}_h \sim \chi^2_l$ hence the $1-\alpha$ quantile of K_h is $q_{1-\alpha}$ the quantile $1-\alpha$ of

a χ^2_l. Then introduce $\text{KICM}_{hC} = \text{KICM}_h \, 1_{G_S \in C_S, G_T \in C_T, G_z \in C_z}$ and $\text{K}_{hC} = \text{K}_h \, 1_{G_\epsilon \in C_S, G_T \in C_T, G_z \in C_z}$. I define the quantile of K_{hC} as $q_{C,1-\alpha} = inf\{q : \mathbb{P}(\text{K}_{hC} \leqslant q) \geqslant 1 - \alpha\}$.

Thus, because $\text{KICM}_h = \text{KICM}_{hC}$ if $G_S \in C_S, G_T \in C_T, G_z \in C_z$ it follows that:

$$\begin{aligned}
1_{\text{KICM}_h > x} &= 1_{\text{KICM}_h > x, G_S \in C_S, G_T \in C_T, G_z \in C_z} + 1_{\text{KICM}_h > x, \overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}} \\
&= 1_{\text{KICM}_{hC} > x, G_S \in C_S, G_T \in C_T, G_z \in C_z} + 1_{\text{KICM}_h > x, \overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}} \\
&\leqslant 1_{\text{KICM}_{hC} > x} + 1_{\overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}}
\end{aligned}$$

By taking the mean it implies that,

$$\mathbb{P}(\text{KICM}_h > x) \leqslant \mathbb{P}(\text{KICM}_{hC} > x) + \mathbb{P}(\overline{G_S \in C_S, G_T \in C_T, G_z \in C_z}) < \mathbb{P}(\text{KICM}_{hC} > x) + \delta$$

Then using $\mathbb{P}(\text{KICM}_h > q_{1-\alpha}) > \alpha + 2\delta$ I obtain:

$$\mathbb{P}(\text{KICM}_{hC} > q_{1-\alpha}) > \alpha + \delta$$

Additionally $\text{KICM}_{hC} \leqslant \text{KICM}_h$ implies that $q_{C,1-\alpha} \leqslant q_{1-\alpha} \forall \alpha$ which leads to,

$$\mathbb{P}(\text{KICM}_{hC} > q_{C,1-\alpha}) > \alpha + \delta$$

Going forward, note that $x \in \mathbb{R}^+ \mapsto x 1_{x<C}$ is bounded by C and Lipschitz. At the same time KICM_h is Lipschitz in G_S, G_T and G_z as long as these take values in bounded sets as saw in B.6 thus KICM_{hC} is Lipschitz in G_S, G_T and G_z. Now that the difference between G_S and G_S^* and between G_T and G_T^* vanishes uniformly by B.4, hence by the Portemanteau Lemma, as KICM_{hC} is Lipschitz and bounded, then KICM_{hC} and K_{hC} converge uniformly towards the same distribution which is the distribution of K_{hC}. It implies by definition of weak convergence that,

$$\sup_x |\mathbb{P}(\text{KICM}_{hC} > x) - \mathbb{P}(\text{K}_{hC} > x)| \to 0$$

Consequently, asymptotically

$$\lim_{n \to \infty} \mathbb{P}(\text{KICM}_{hC} > q_{C,-\alpha}) = \mathbb{P}(\text{K}_{hC} > q_{C,1-\alpha}) = \alpha \geqslant \alpha + \delta \Leftrightarrow \delta \leqslant 0$$

Which is impossible because $\delta > 0$.

On a final note, the Lipschitzness and boundedness of KICM_{hC} and K_{hC} in G_S and G_T imply that the difference between the feasible and unfeasible statistics vanishes uniformly by asymptotic equicontinuity uniform of the processes G_S and G_T. Therefore the contradiction still holds even if KICM_{hf} is used instead of KICM_h, and K_{hf} instead of K_h (where K_{hf} denotes KICM_{hf} but with Y_i conditionally normal).

C Proof of Corollary 4.4

For any null β_0, G_T is still bounded and bounded away from 0 as shown in B.4 uniformly over the $\mathbb{P} \in \mathcal{P} : \beta \neq \beta_0$. G_T's limit is different under the alternative tough as it depends on the actual distribution of y_i. As for G_S and G_S^* the difference between the two will also vanish as per the arguments of B.4, even if G_S and G_S^* don't have a limit but explode. Before deriving the power of KICM under different types of identification let's first decompose G_S under the alternative $\beta \neq \beta_0$:

$$G_S(t,\Omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ih}e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{y_i - x_i'\beta_0}{\sqrt{b_0'\Omega_i b_0}}e^{it'z_i} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{y_i - x_i'\beta}{\sqrt{b_0'\Omega_i b_0}}e^{it'z_i} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i'(\beta-\beta_0)}{\sqrt{b_0'\Omega_i b_0}}e^{it'z_i}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(u_i + v_i'(\beta-\beta_0))\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}} + \frac{1}{n^{1/2+a}}\sum_{i=1}^{n} C(z_i)'(\beta-\beta_0)\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}}$$

The term on the left will, reusing arguments from B.4, converge towards a complex Gaussian process with mean 0 and covariance $(s,t) \mapsto \mathbb{E}(Me^{i(s+t)'z_i})$ with

$$M = (1\ \beta' - \beta_0')\frac{\Omega_i}{b_0'\Omega_i b_0}(1\ \beta' - \beta_0')'$$

As a consequence

$$G_S(t,\Omega) = O_{\bar{\mathbb{P}}}(1) + \frac{n^{1/2-a}}{n}\sum_{i=1}^{n} C(z_i)'(\beta-\beta_0)\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}} = O_{\bar{\mathbb{P}}}(1) + O_{\bar{P}}(n^{1/2-a})$$

And there are 3 possible behaviors of KICM:

1. If $a < 1/2$ then $\frac{G_S(t,\Omega)}{n^{1/2-a}} = O_{\bar{\mathbb{P}}}(1) \neq o_{\underline{\mathbb{P}}}(1)$. In other words $G_S(\cdot,\Omega)/n^{1/2-a}$ is uniformly bounded in probability and uniformly non-degenerate and converges towards the following function,

$$t \mapsto \mathbb{E}(C(z_i)'\frac{e^{it'z_i}}{\sqrt{b_0'\Omega_i b_0}})(\beta-\beta_0)$$

   Then because $G_S/n^{1/2-a}$, G_T and G_z are $O_{\bar{\mathbb{P}}}(1)$ then for any $\zeta > 0$ there exists bounded subsets $C_S \subset \mathbb{C}_*$, $C_T \subset \mathbb{C}_*^l$ and $C_z \in \mathbb{C}_*$ such that $P(G_S/n^{1/2-a} \notin C_S, G_T \notin C_T, G_z \notin C_z) < \zeta$. Also note that, conditionally on the event $C = (G_S/n^{1/2-a} \in C_S, G_T \in C_T, G_z \in C_z)$, $\text{KICM}_h\ /n^{1-2a} > K$ surely where K is a strictly positive constant (by properties of integrals).

Now assume that for some $x \in \mathbb{R}^+$, $\mathbb{P}(\text{KICM}_h > x) \to 1$ doesn't hold. Then it implies that $\exists \eta \in [0;1)$ such that $\lim_{n\to\infty} \mathbb{P}(\text{KICM}_h > x) < \eta$ or equivalently that $\exists(\eta, \zeta) \in [0;1) \times \mathbb{R}_*^+$ such that $\lim_{n\to\infty} \mathbb{P}(\text{KICM}_h > x) \leqslant \eta - \zeta$. Additionally,

$$\begin{aligned}
\eta - \zeta \geqslant \mathbb{P}(\text{KICM}_h > x) &= \mathbb{P}(\text{KICM}_h > x|C)\mathbb{P}(C) + \mathbb{P}(\text{KICM}_h > x|\bar{C})\mathbb{P}(\bar{C}) \\
&\geqslant \mathbb{P}(\text{KICM}_h > x|C)\mathbb{P}(C) \\
&= \mathbb{P}(\text{KICM}_h / n^{1-2a} > x/n^{1-2a}|C)(1-\zeta) \\
&\geqslant \mathbb{P}(K > x/n^{1-2a}|C)(1-\zeta) \\
&= \mathbb{P}(K > x/n^{1-2a})(1-\zeta) \\
&\to 1 - \zeta
\end{aligned}$$

Which is equivalent to $\eta - \zeta \geqslant 1 - \zeta$ asymptotically which is impossible for any $x \in \mathbb{R}^+$ thus the test is consistent. It trivially follows that this contradiction holds uniformly over the β_0 and the $P : \beta \neq \beta_0,\ a < 1/2$ per the results in B.3.

2. If $a = 1/2$ then G_S and G_S^* converges towards the same non-centered complex Gaussian process. This means that the distance between KICM_h under conditional normality and KICM_h still vanishes. Then KICM_h does not follow a χ^2_l asymptotically anymore but a non centered $\chi^2_{l,\lambda}$ where $\lambda = \mathbb{E}\left(\frac{C(z_i)'(\beta-\beta_0)}{\sqrt{b_0'\Omega_i b_0}}\right)^2$. Indeed when Y_i is conditionally normal then even if $\mathbb{E}(S_{ih}|z_i) \neq 0$, $\text{Var}(S_{ih}|z_i) = 1$ and $\text{Cov}(S_{ih}, T_{ih}) = 0_l$ which implies that S is independent from T. Consequently, under conditional normality KICM_h is indeed the sum of l independent non-centered standard normal variable thus conditionally on z and T and assuming conditional normality $\text{KICM}_h \sim \chi^2_{l,\lambda}$. Furthermore, the cdf of a non-central χ^2 evaluated at any point is strictly decreasing in its non-centrality parameter. Therefore,

$$\begin{aligned}
&\mathbb{P}(\text{KICM}_h > x|z, T) \to \mathbb{P}(\chi^2_{l,\lambda} > x) > \mathbb{P}(\chi^2_l > x) \\
\Rightarrow &\lim_{n\to\infty} \mathbb{P}(\text{KICM}_h > q_{1-\alpha}) = \lim_{n\to\infty} \mathbb{E}(\mathbb{P}(\text{KICM}_h > q_{1-\alpha}|z, T)) > \alpha
\end{aligned}$$

In other words, there is more than trivial power. As the convergence towards a $\chi^2_{l,\lambda}$ can be proven to be uniform over the β_0 and $\mathbb{P} : \beta \neq \beta_0,\ a = 1/2$ per arguments in B.3

$$\lim_{n\to\infty} \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0,\ a=1/2} \mathbb{P}(\text{KICM}_h > q_{1-\alpha}) \geqslant \alpha$$

3. If $a > 1/2$ then reusing arguments from B.4 the processes $G_S(\cdot, \Omega)$ and $G_S^*(\cdot, \Omega)$ will converge to the same centered complex Gaussian process as under the null because S_{ih} has mean 0 asymptotically. Therefore, following the proof of Theorem 4.3, KICM_h

will still behave as in the conditionally normal case. From there, because $\text{Var}(S) = I_n$ still, KICM_h will behave like the sum of l standard normal and follow a χ^2_l. Thus the probability to reject the null at nominal level α will be inferior or equal to α uniformly over the $P : \beta \neq \beta_0$ and the β_0. The omitted part of the proof is a copy of the proof of Theorem 4.3.

D Non-uniform identification strength

It is important to note that one could assume that the parameters do not have the same degree of identification as in assumption B(iii) instead of B(ii). Then the asymptotic behavior of the KICM test would remain relatively unchanged using a new a_j^* to characterize the asymptotics

$$a_j^* = \begin{cases} min\{a_j : \beta_j \neq \beta_{0j}\} & \text{if}(w^*_{ij,l'})_{l'=1}^{l} \neq 0_l \\ 1 & \text{o.w} \end{cases}$$

Non-uniformity of the degree of identification of β does not affect the size of KICM, but affects its power. Among the components j of β such that the null being tested $\beta_{0j} \neq \beta_j$ if at least one parameter is semi-strongly identified then the test is consistent, if at least one parameter is weakly identified then the test has more than trivial power, if all the parameters are very weakly identified then the test has trivial power. These results are summarized in the following corollary

Corollary 4.1 (Uniform consistency of KICM, non-uniform identification strength)
Denote by $q_{1-\alpha}$ the $1-\alpha$ quantile of the chi-square distribution with l degrees of freedom. Then under assumptions A(ii), A(iii), B(i), B(iii), B(iv), C(i), C(ii), D(ii), and D(iii),

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}<1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) = 1;$
  The test is consistent when at least one parameter is semi-strongly identified.

- $\lim_{n\to\infty} \inf_{\beta_0} \inf_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}=1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) \in [\alpha; 1);$
  The test has more than trivial power when at least one parameter is weakly identified.

- $\lim_{n\to\infty} \sup_{\beta_0} \sup_{\mathbb{P}\in\mathcal{P}:\beta\neq\beta_0, min\{a_j:\beta_j\neq\beta_{0j}\}>1/2} \mathbb{P}(\text{KICM}_{hf} > q_{1-\alpha}) \leqslant \alpha;$
  The test has trivial power when all parameters are very weakly identified.

The proof is extremely similar to that of Corollary 4.4 in appendix C and is omitted.

E Plots and tables

E.1 Small sample simulations

E.1.1 Empirical size

Weak instruments	AR	LM	CLR	ICM	KICM	CICM	W-2SLS
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1112	0.1398	0.1684
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1112	0.1122	0.1684
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.1004	0.0982	0.1592
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.1004	0.1074	0.1592
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1070	0.1124	0.1848
Non-linear model							
n=100, m=200	0.1170	0.1120	0.1174	0.1364	0.1072	0.1354	0.3964
n=100, m=500	0.1170	0.1120	0.1158	0.1176	0.1072	0.1180	0.3964
n=400, m=200	0.0980	0.0958	0.0942	0.1052	0.0946	0.0992	0.3878
n=400, m=500	0.0980	0.0958	0.0944	0.1122	0.0946	0.1056	0.3878
n=400, m=500, Heteroskedasticity	0.1044	0.1044	0.1044	0.1174	0.1014	0.1060	0.3856
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1034	0.1378	0.2030
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1034	0.1200	0.2030
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0976	0.1006	0.1986
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0976	0.1090	0.1986
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1028	0.1174	0.1992
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1092	0.1134	0.1364	0.1106	0.1366	0.3908
n=100, m=500	0.1170	0.1092	0.1108	0.1176	0.1106	0.1176	0.3908
n=400, m=200	0.0980	0.0958	0.0948	0.1052	0.0954	0.0998	0.3848
n=400, m=500	0.0980	0.0958	0.0934	0.1122	0.0954	0.1060	0.3848
n=400, m=500, Heteroskedasticity	0.1048	0.1010	0.1040	0.1186	0.0990	0.1084	0.3786

Table 1: Empirical size of the tests for nominal size 10%, weak instruments case

Semi-strong instruments	AR	LM	CLR	ICM	KICM	CICM	W-2SLS
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1098	0.1238	0.1078
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1098	0.1174	0.1078
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0984	0.0998	0.0908
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0984	0.0996	0.0908
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1038	0.1060	0.1024
Non-linear model							
n=100, m=200	0.1170	0.1152	0.1128	0.1364	0.894	0.1228	0.3206
n=100, m=500	0.1170	0.1152	0.1134	0.1176	0.894	0.1106	0.3206
n=400, m=200	0.0980	0.1000	0.0948	0.1052	0.0902	0.0936	0.2844
n=400, m=500	0.0980	0.1000	0.0956	0.1122	0.0902	0.1014	0.2844
n=400, m=500, Heteroskedasticity	0.1048	0.1016	0.1014	0.1164	0.0980	0.1028	0.2940
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.0998	0.1350	0.1788
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.0998	0.1152	0.1788
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0994	0.0990	0.1764
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0994	0.1056	0.1764
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1150	0.1042	0.1086	0.1758
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1102	0.1106	0.1364	0.0950	0.1200	0.2620
n=100, m=500	0.1170	0.1102	0.1088	0.1176	0.0950	0.1098	0.2620
n=400, m=200	0.0980	0.0936	0.0942	0.1052	0.0918	0.0976	0.1972
n=400, m=500	0.0980	0.0936	0.0958	0.1122	0.0918	0.1010	0.1972
n=400, m=500, Heteroskedasticity	0.1048	0.1042	0.1018	0.1186	0.0968	0.1016	0.2136

Table 2: Empirical size of the tests for nominal size 10%, semi-strong instruments case

Strong instruments	AR	LM	CLR	ICM	KICM	CICM	W-2SLS
Linear model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.1124	0.1128	0.1008
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.1124	0.1118	0.1008
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.0976	0.0938	0.0954
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.0976	0.0962	0.0954
n=400, m=500, Heteroskedasticity	0.1056	0.1056	0.1050	0.1164	0.1068	0.1086	0.0986
Non-linear model							
n=100, m=200	0.1070	0.1056	0.1064	0.1364	0.0914	0.1120	0.1478
n=100, m=500	0.1070	0.1056	0.1030	0.1176	0.0914	0.0964	0.1478
n=400, m=200	0.0980	0.0968	0.0962	0.1052	0.966	0.1012	0.1040
n=400, m=500	0.0980	0.0968	0.0944	0.1122	0.966	0.1018	0.1040
n=400, m=500, Heteroskedasticity	0.1048	0.0960	0.0948	0.1122	0.0964	0.1046	0.0934
Polar polynomial model							
n=100, m=200	0.1042	0.1042	0.1044	0.1414	0.0958	0.1114	0.0836
n=100, m=500	0.1042	0.1042	0.1028	0.1188	0.0958	0.1036	0.0836
n=400, m=200	0.0976	0.0976	0.0944	0.1024	0.1000	0.1004	0.0708
n=400, m=500	0.0976	0.0976	0.0974	0.1096	0.1000	0.1032	0.0708
n=400, m=500, Heteroskedasticity	0.1050	0.1050	0.1036	0.1074	0.0980	0.1052	0.0404
Semi-polar polynomial model							
n=100, m=200	0.1170	0.1096	0.1094	0.1364	0.0902	0.0996	0.1156
n=100, m=500	0.1170	0.1096	0.1074	0.1176	0.0902	0.0924	0.1156
n=400, m=200	0.0980	0.0974	0.0940	0.1052	0.0944	0.1006	0.0960
n=400, m=500	0.0980	0.0974	0.0962	0.1122	0.0944	0.0984	0.0960
n=400, m=500, Heteroskedasticity	0.08	0.10	0.10	0.07	0.10	0.10	0.06

Table 3: Empirical size of the tests for nominal size 10%, strong instruments case

Instruments Strength	**AR**	**LM**	**CLR**	**ICM**	**KICM**	**CICM**	**W-2SLS**
Weak	0.1298	0.1186	0.1270	0.1438	0.0900	0.1370	0.7052
Semi-Strong	0.1298	0.1178	0.1232	0.1438	0.0868	0.1362	0.6506
Strong	0.1298	0.1078	0.1110	0.1438	0.0984	0.1250	0.3704

Table 4: Empirical size for nominal size 10%, 4 Instruments

E.1.2 Power curves

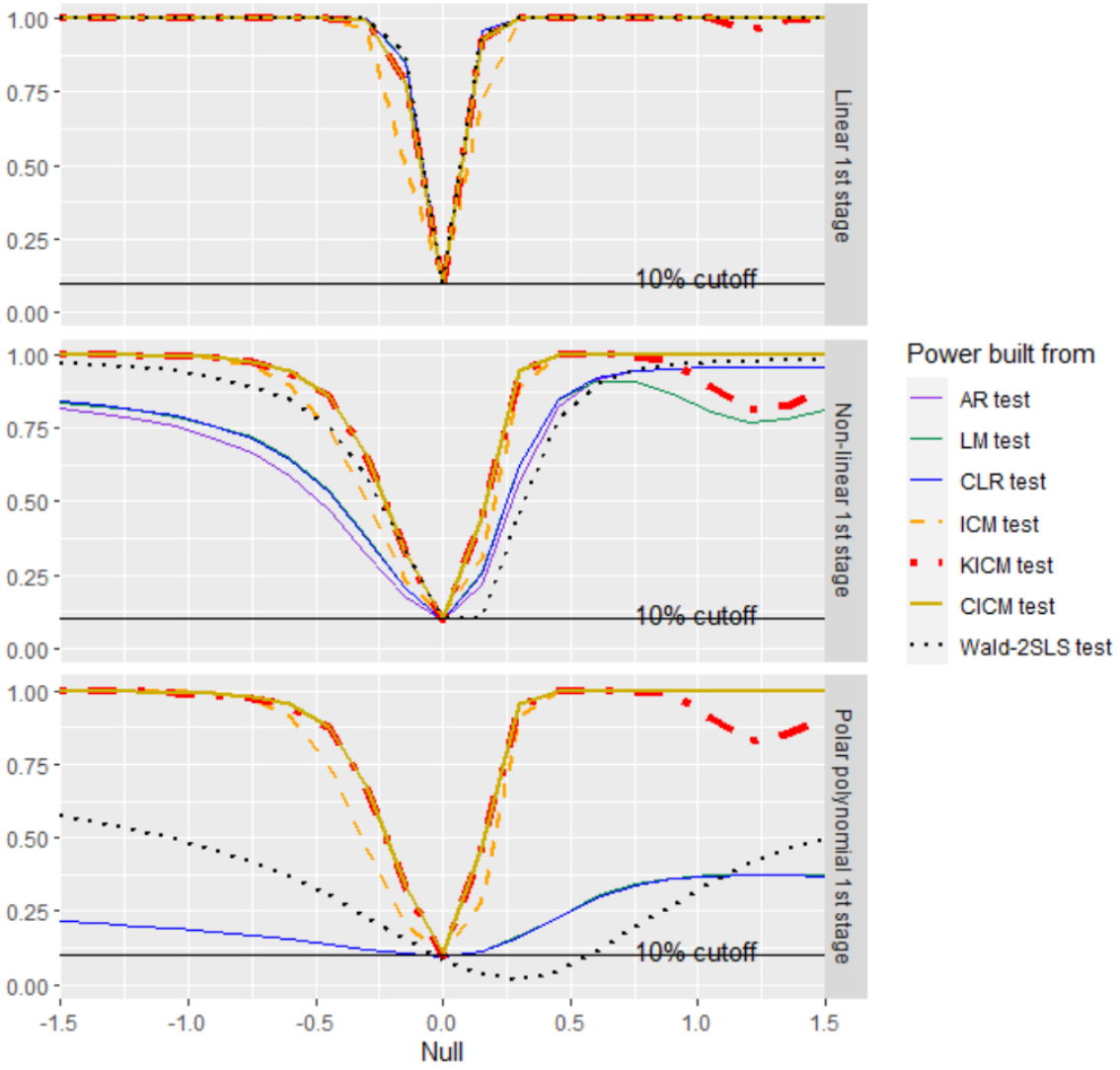


Figure 1: Power curves, strong instruments, homoskedastic Data

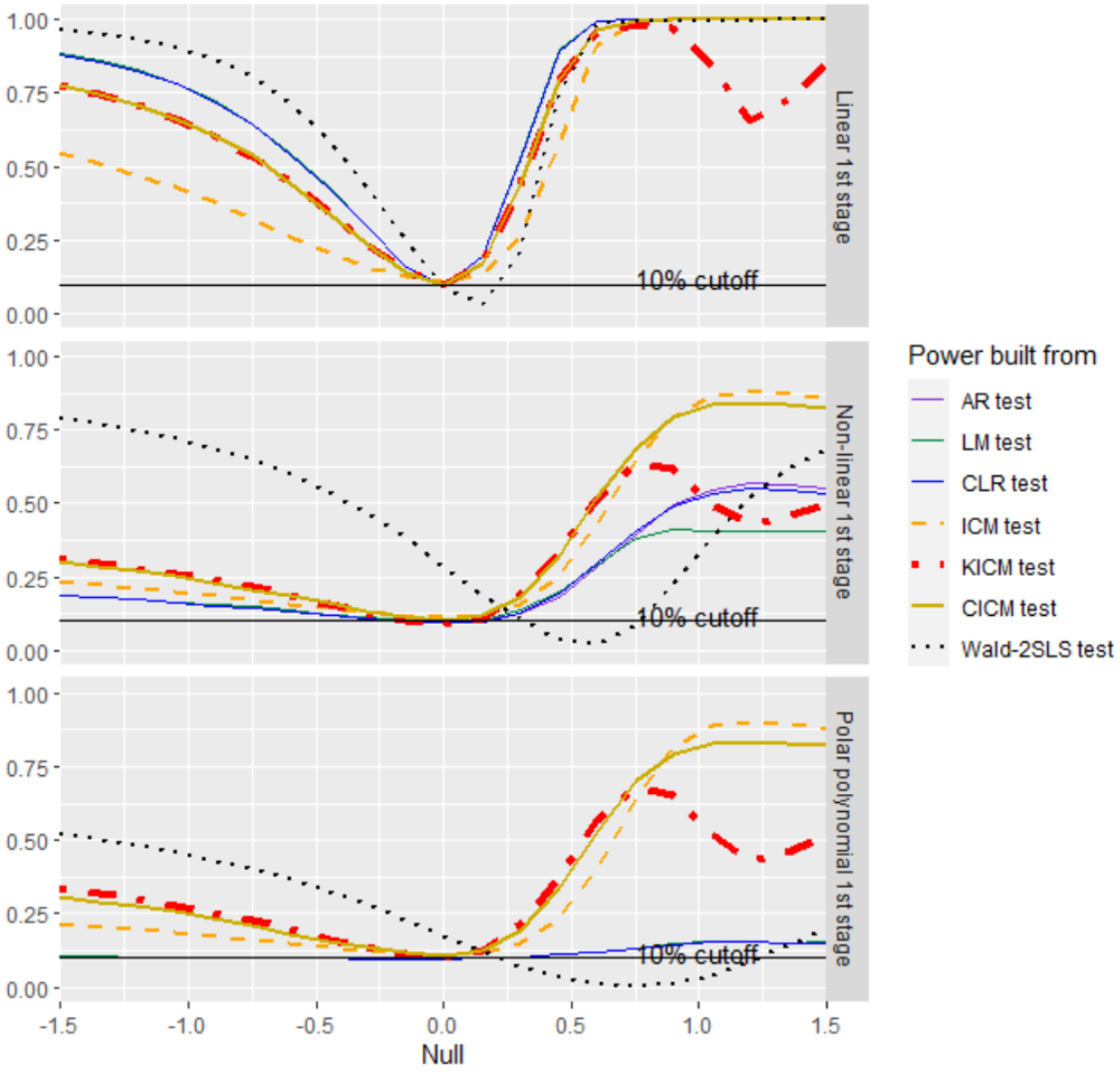


Figure 2: Power curves, semi-strong instruments, homoskedastic data

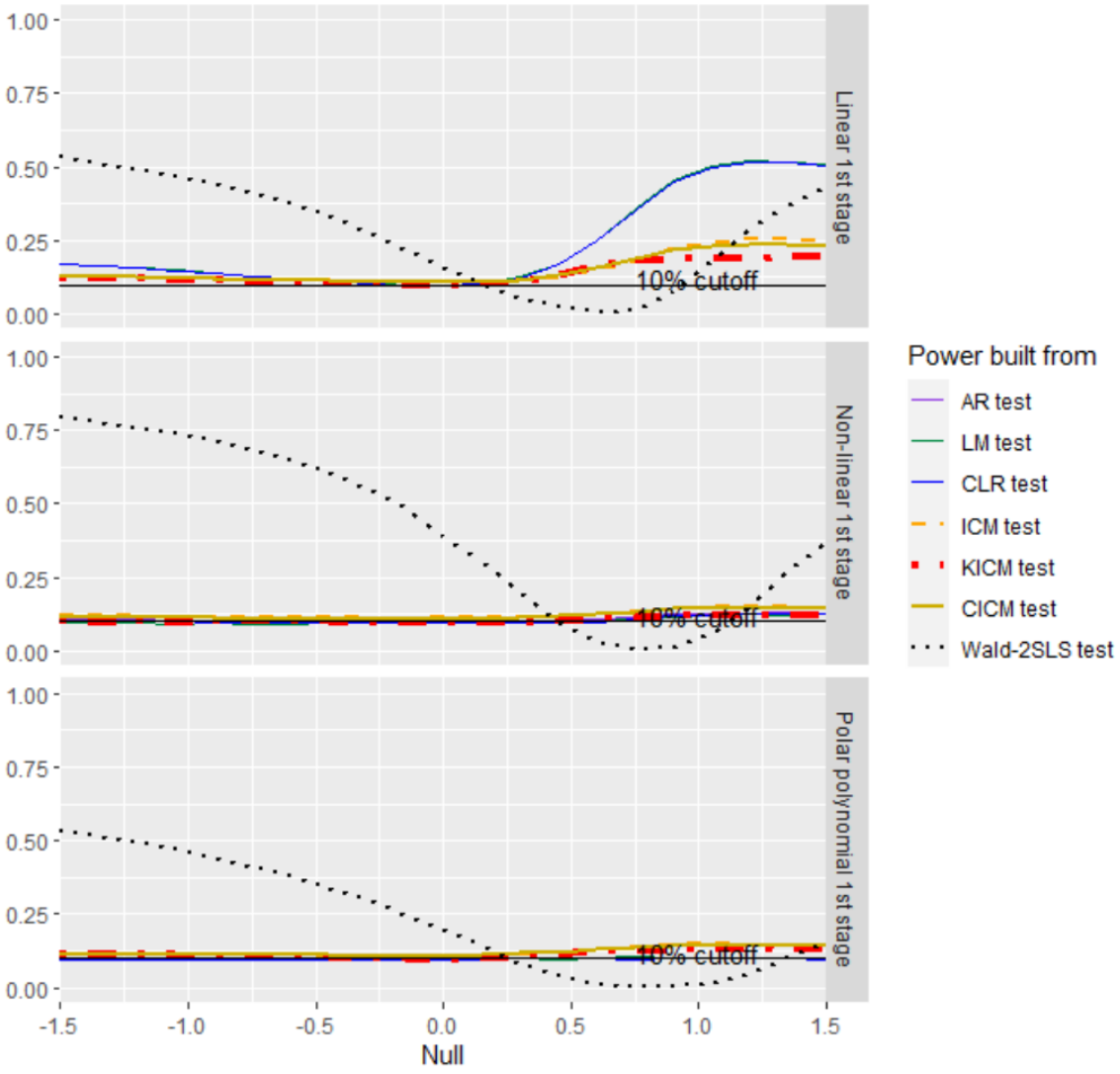


Figure 3: Power curves, weak instruments, homoskedastic data

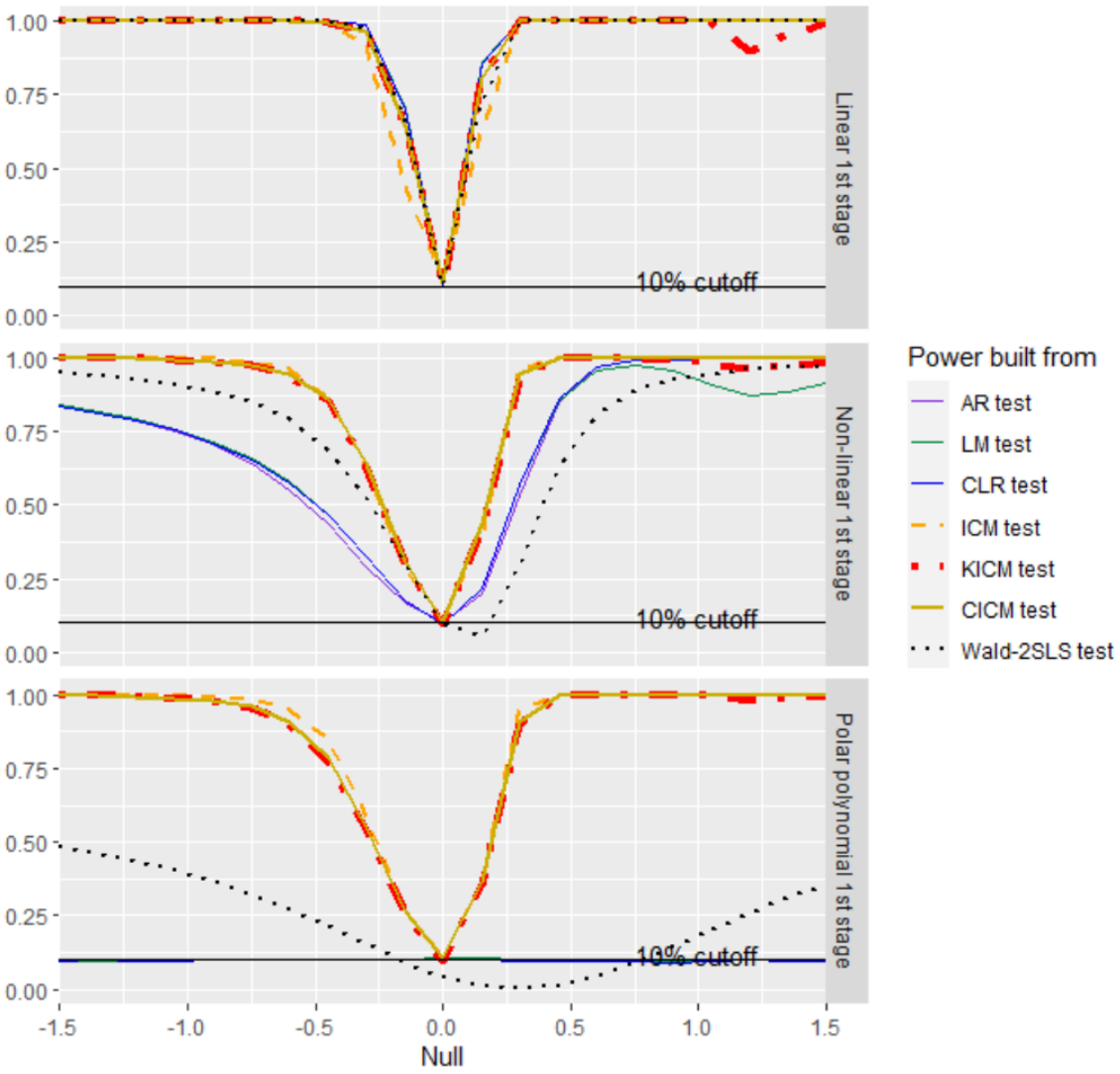


Figure 4: Power curves, strong instruments, heteroskedastic data

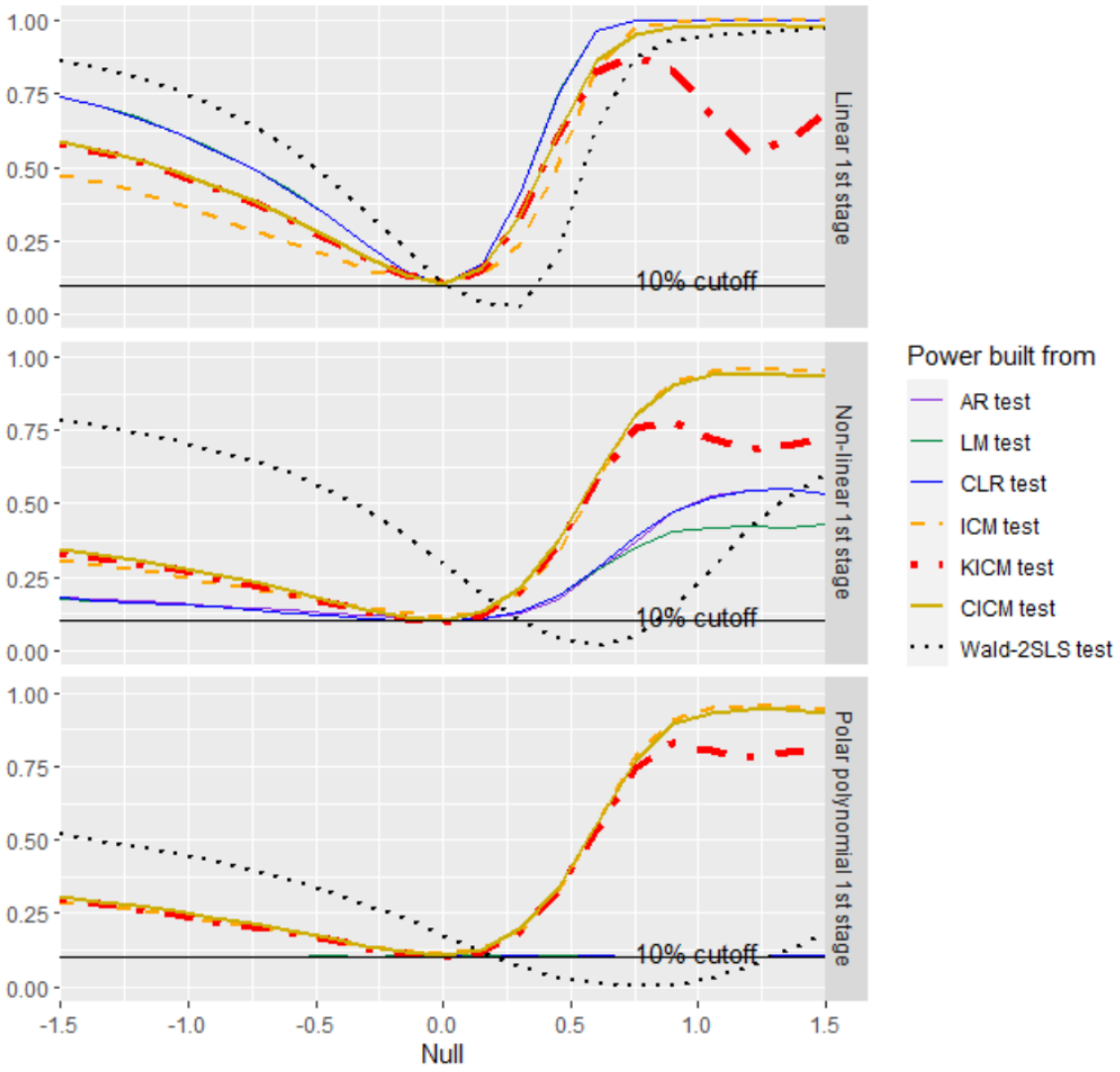


Figure 5: Power curves, semi-strong instruments, heteroskedastic data

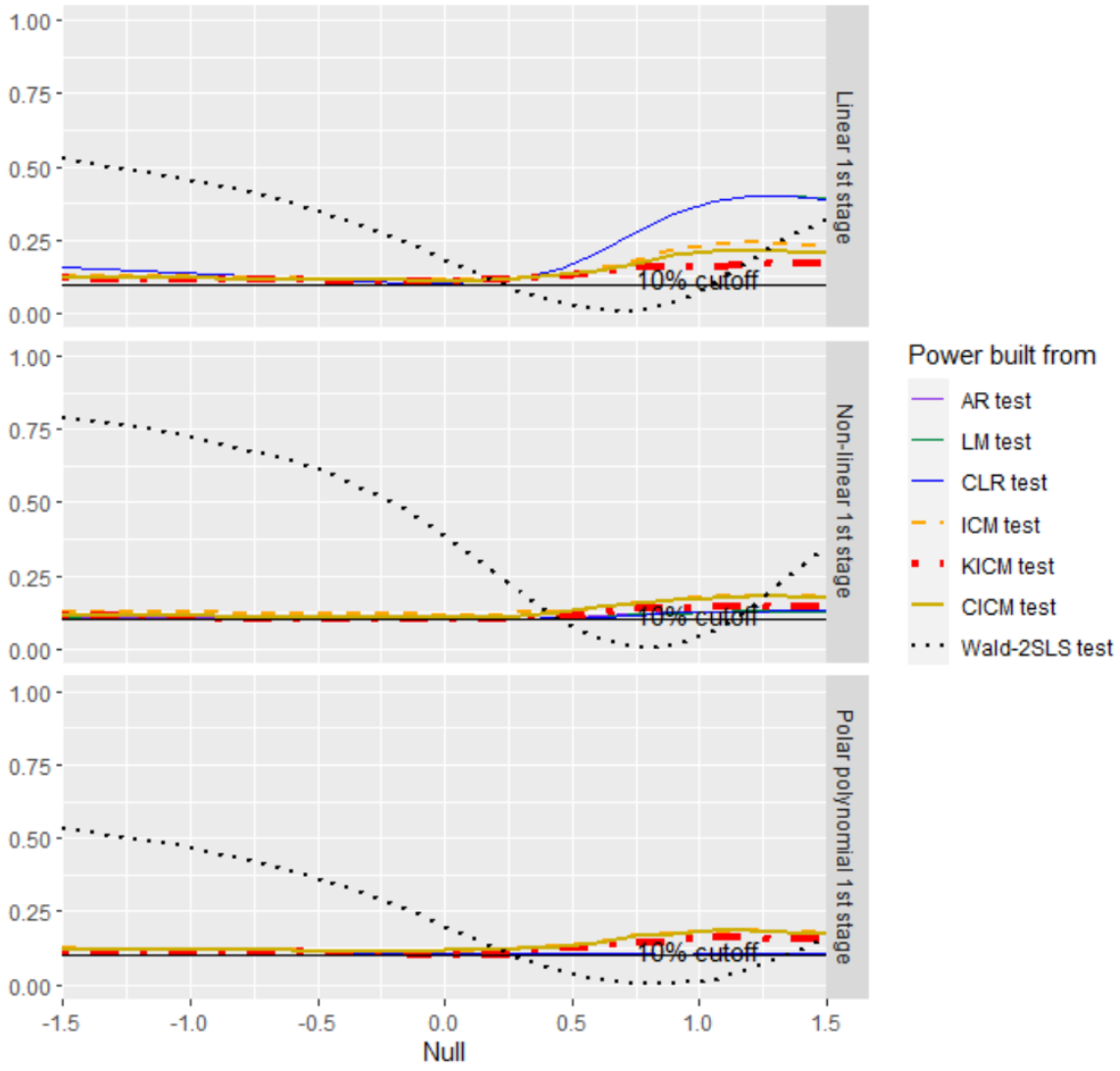


Figure 6: Power curves, weak instruments, heteroskedastic data

E.1.3 Average p-value curves

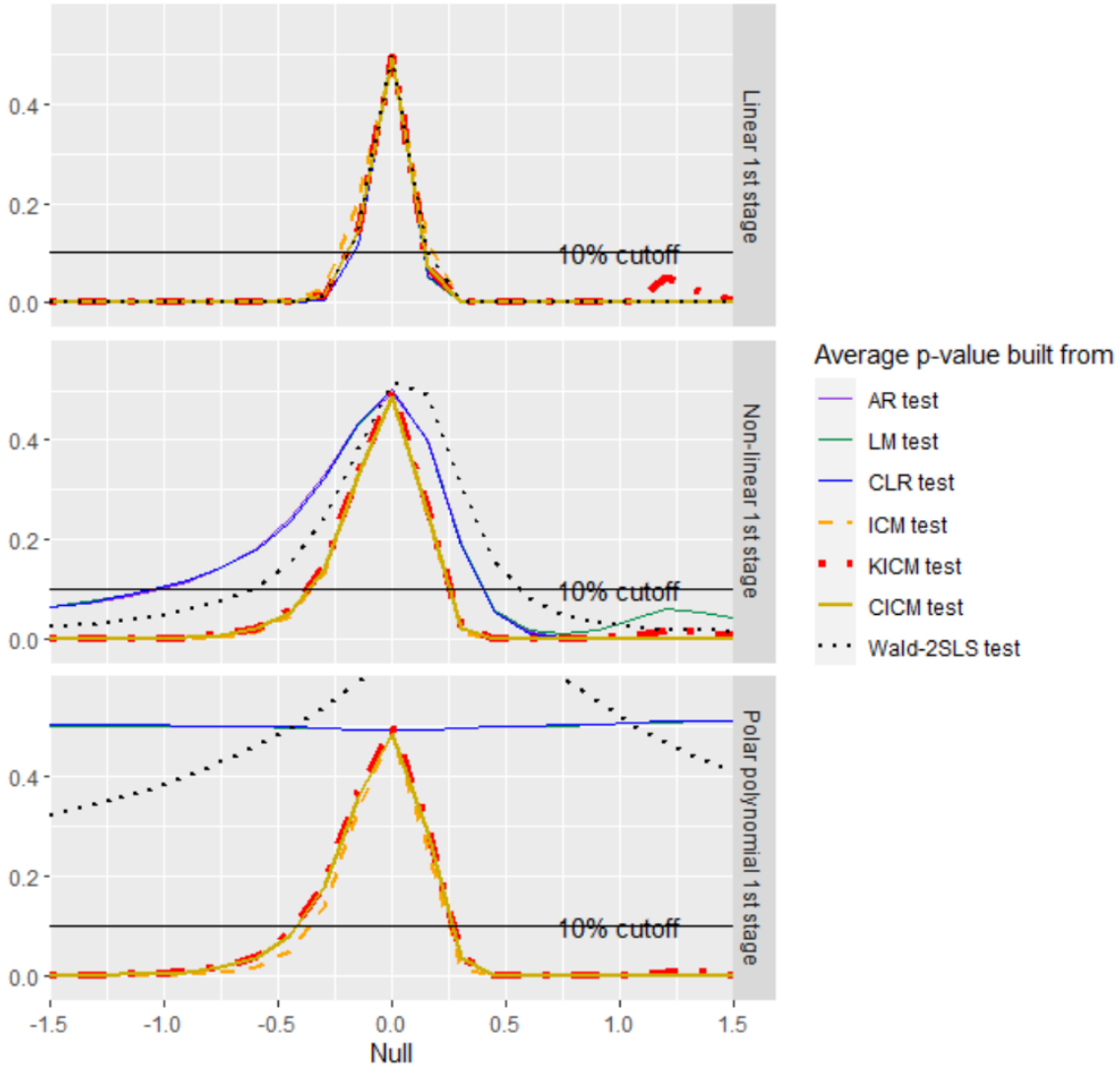


Figure 7: P-value curves, strong instruments, heteroskedastic data

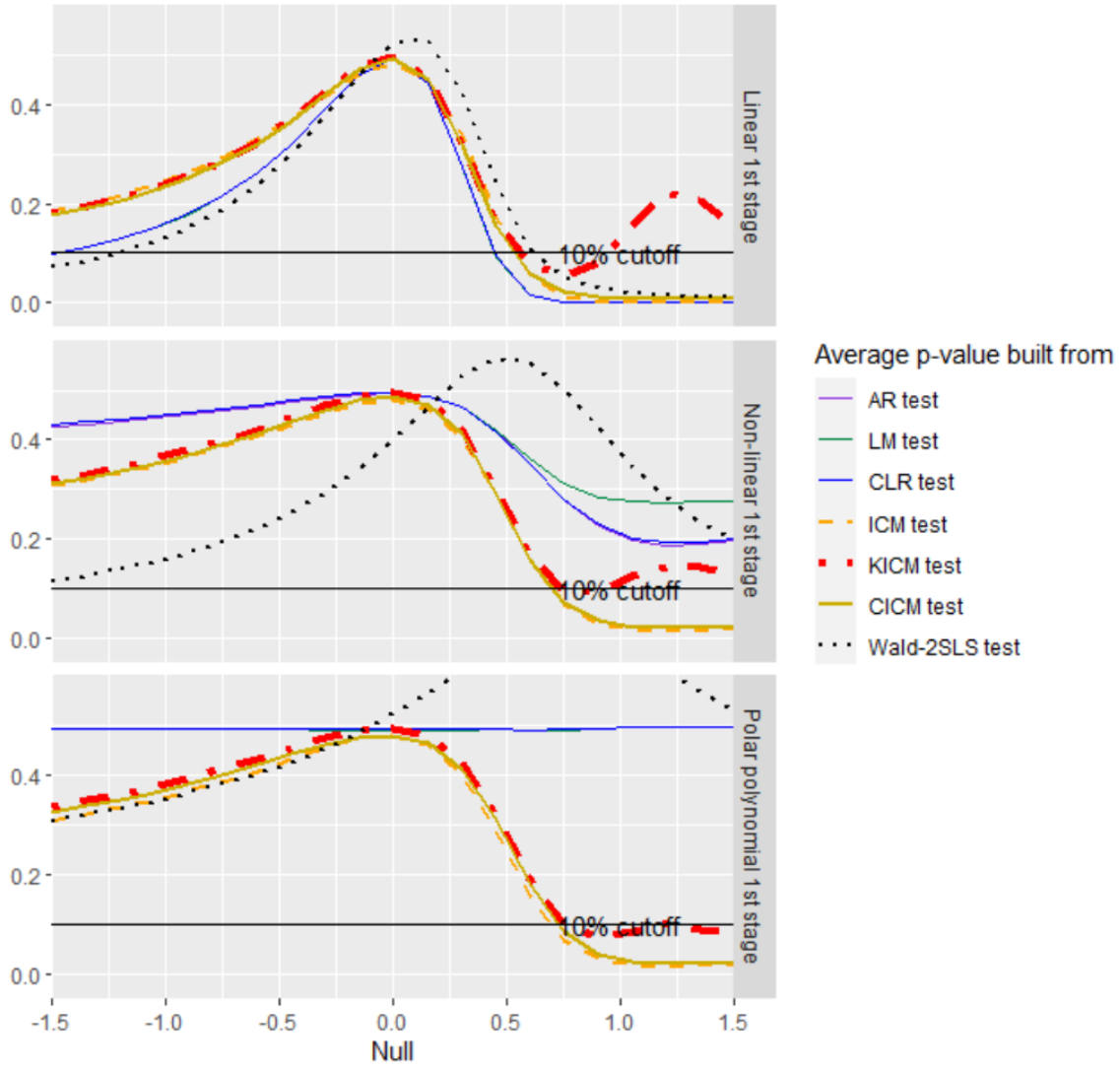


Figure 8: P-value curves, semi-strong instruments, heteroskedastic data

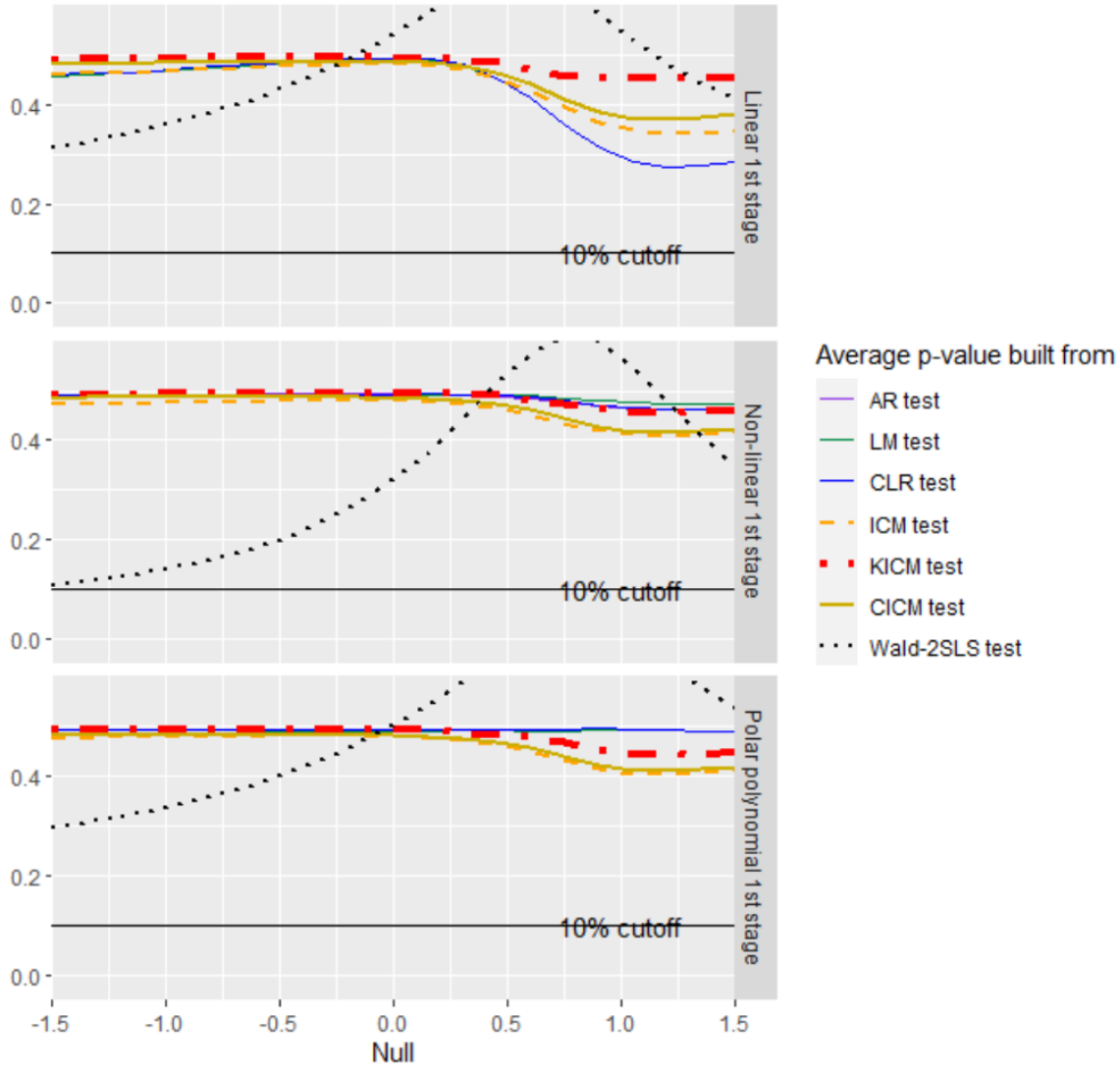


Figure 9: P-value curves, weak instruments, heteroskedastic data

E.2 Application

Specification	(1)	(2)	(3)	(4)
KICM	[0.009;0.132]	$\mathbb{R}$	$\mathbb{R}$	[0.138;0.240]
AR	[0.009;0.162]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
LM	[0.047;0.118]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
CLR	[0.042;0.110]	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
OLS	0.080	0.080	0.072	0.070
	[0.080;0.081]	[0.080;0.081]	[0.071;0.072]	[0.070;0.071]
2SLS	0.077	0.131	0.106	0.101
	[0.052;0.102]	[0.076;0.186]	[0.050;0.163]	[0.046;0.156]
LIML	0.076	0.255	0.300	0.282
	[0.047;0.105]	[0.118;0.393]	[0.068;0.531]	[0.059;0.505]
FULLER	0.76	0.238	0.256	0.241
	[0.047;0.105]	[0.100;0.375]	[0.024;0.487]	[0.018;0.464]
Age and age square	-	Yes	Yes	Yes
Additional covariates	-	-	Yes	Yes
Region residence FE	-	-	-	Yes
First, stage F test statistic	4.68	1.08	0.99	1.03

Table 5: 90% confidence intervals for returns to education, cohort 20-29

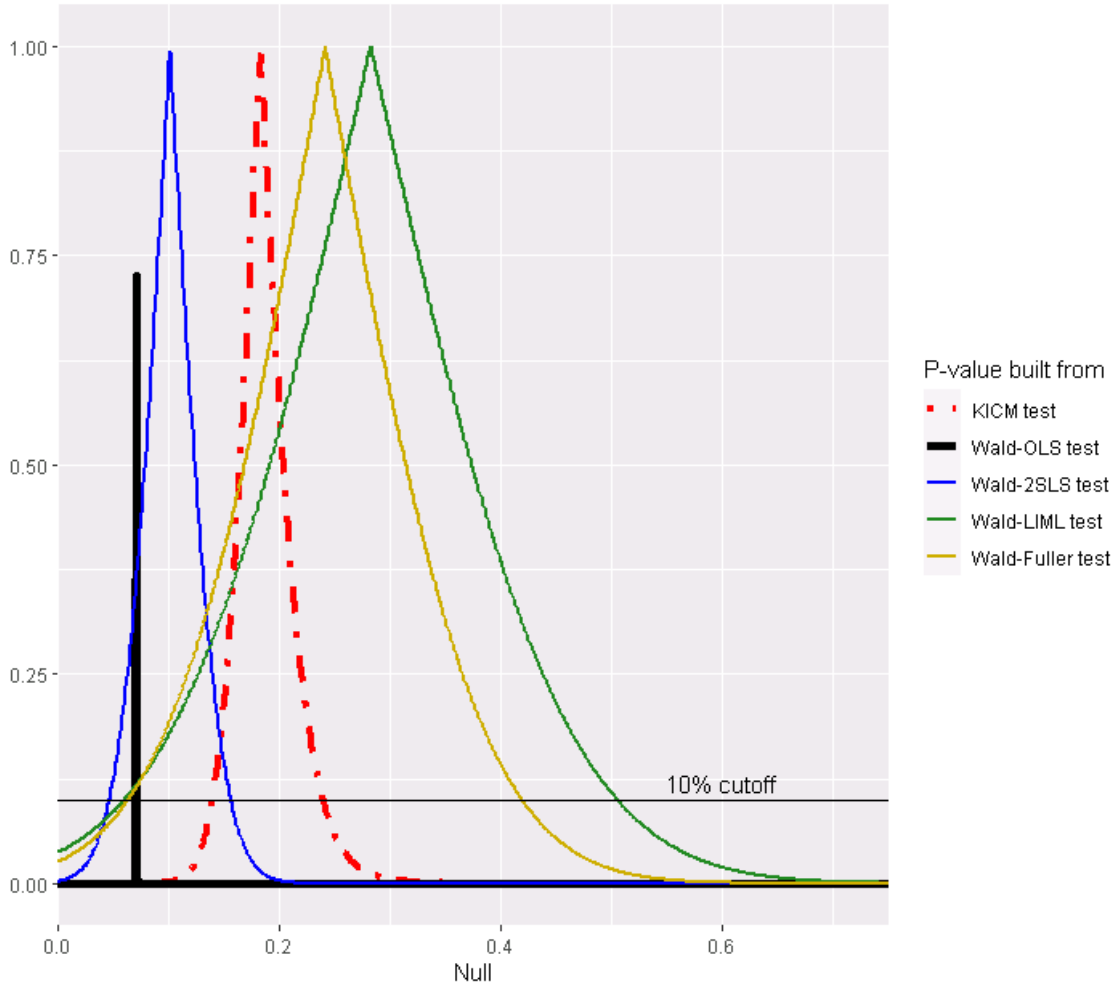


Figure 10: P-value curve of return to education, all covariates and fixed effects setting, cohort 20-29

Chapter 2: Testing and Relaxing Distributional Assumptions on Random Coefficients in Demand Models

Co-authored with Gökçe Gökkoca and Max Lesellier

Abstract

The BLP demand model for differentiated products is the workhorse model for demand estimation with market-level data. This model uses random coefficients to account for unobserved preference heterogeneity. The shape of the distribution of random coefficients matters greatly for many counterfactual quantities, such as the cost pass-through. In this paper, we develop new econometric tools to test this distribution and improve its estimation under a flexible parametrization. First, we construct new instruments that are designed to detect deviations from the true distribution of random coefficients. Second, we develop a formal moment-based specification test on the distribution of random coefficients. Third, we show that our instruments can be successfully used to estimate a flexible distribution of random coefficients. Finally, we validate our approach with Monte Carlo simulations and an empirical application using data on car purchases in Germany. We also show that these methods extend to the mixed logit demand model with individual-level data.

Keywords: Demand Estimation, Specification Test, Random Coefficients
JEL codes: C35, C36, L13, C52

1 Introduction

The differentiated product demand model initiated by Berry (1994) and Berry, Levinsohn, and Pakes (1995) has been used in a wide array of empirical studies. It enables researchers to perform demand estimation in markets with differentiated products using either macro-level (market shares) or micro-level (individual purchases) data while allowing for unobserved heterogeneity in preferences as well as price endogeneity. This unobserved heterogeneity in preferences is modeled through the use of random coefficients (RCs) in the utility function. This framework allows researchers to estimate demand functions, price elasticities and counterfactual outcomes. Applications of the BLP model have notably studied the determinants of market power, the welfare effects resulting from a merger or the introduction of a new good and the economic impact of a tax or a subsidy.[1]

The informativeness of the empirical analysis depends on how well the model can reproduce the underlying substitution patterns and approximate the shape of the demand curve, including its slope and curvature. A recent result in Miravete, Seim, and Thurk (2022) shows that the commonly used Gaussian RC on price imposes strong restrictions on the demand's curvature and thus limits the range of the implied pass-through. The degree of pass-through of taxes and costs is central to answering many questions in economics such as the impact of tariffs or a cost shock on consumer welfare. However, estimating a more flexible demand system with a non-Gaussian distribution of random coefficients is challenging. First, there is a clear trade-off between the degree of flexibility one chooses (for instance, going from a Gaussian to a Gaussian mixture) and the precision of the estimates one obtains. Therefore, it is important to be able to test the specification chosen by the researcher on the distribution of the RC (for instance, a Gaussian RC) and quantify the degree of misspecification before potentially moving to a more flexible specification. Second, to precisely estimate a more flexible distribution of RC, the researcher must choose instruments (or equivalently moment conditions) that strongly identify this distribution. The instruments used by the current empirical practice work well with the standard Gaussian RC, but their performance appears to decline as the specification becomes more flexible in the simulation exercises that we perform.

[1]The BLP demand model has been widely utilized in numerous applications. A non-exhaustive list of examples includes: Barahona, Otero, Otero, and Kim (2020), Berry et al. (1995), Crawford, Shcherbakov, and Shum (2019), Dubois, Griffith, and O'Connell (2018), Durrmeyer (2022), Grennan (2013), Grigolon, Reynaert, and Verboven (2018), Miller, Sheu, and Weinberg (2021), Miller and Weinberg (2017), Miravete, Moral, and Thurk (2018), Nevo (2000), Petrin (2002), Reynaert (2021).

In this paper, we provide novel econometric tools to address these two challenges. In particular, we construct a new set of instruments designed to detect deviations from the true distribution of random coefficients. Building on these instruments, we provide a formal moment-based specification test on the distribution of random coefficients, which can be implemented without having to re-estimate the model under a more flexible parametrization. Our instruments are designed to maximize the power of this test when the distribution of RC is misspecified. We also show how these instruments can strengthen the identifying power of the moment conditions used for estimation, and thus be successful at estimating a flexibly parameterized distribution of RCs. As an example of a flexible parametric distribution, we consider the Gaussian mixture, which can approximate arbitrarily well any continuous distribution on the real line.

This paper consists of three main contributions. First, we construct a new set of instruments that are designed to detect departures from the true distribution of RCs. The intuition we use is the following. Any given distribution of RCs generates a structural error, which, if correctly specified, is mean-independent with respect to a set of exogenous variables. This identifying condition can be transformed into unconditional moments, which can be used to test whether the chosen distribution of RCs is correctly specified. We formally define this test and construct instruments that maximize its power against a fixed alternative. In a first step, we assume that the econometrician knows the fixed alternative and derives an expression for the first-best instrument. We call this instrument the most powerful instrument (MPI) and show that this specific choice of instrument achieves the consistency of the test. In a second step, we provide two feasible approximations of the MPI that can be derived without knowledge of the fixed alternative. We call these feasible MPIs the interval instruments in reference to the way they approximate the MPI.

Second, we consider the case where the researcher wants to test whether the distribution of RCs belongs to a given parametric family. For instance, the researcher may be interested in testing if the random coefficient is normally distributed. This is a composite hypothesis, and we must estimate the unknown parameters of the distribution in a first step. In a second step, we choose instruments to test if the distribution evaluated at the estimated parameters is correctly specified. Here, the interval instruments represent a natural choice of instruments as they are designed to detect deviations from the true distribution of RCs. We study the asymptotic properties of our test when the number of markets, T, goes to infinity and we prove the asymptotic validity of the test under common assumptions. In particular, we ac-

count for the statistical uncertainty stemming from the first step estimation, and we control for the magnitude of the approximations that intervene in the estimation of the BLP model. Our asymptotic results complement previous work by Freyberger (2015) on the asymptotic properties of the BLP estimator when the number of markets grows to infinity.

Third, we show that our interval instruments can be successfully used to estimate the model, and particularly so when the distribution of RCs is flexibly parameterized. We do so by exhibiting the connection between the MPI and the classical optimal instruments used for efficient estimation purposes. Specifically, we show that the MPI devoted to testing the specification of the model at the true parameter against any local alternative can be rewritten as a linear combination of the optimal instruments. This relation between the MPI and the optimal instruments helps us understand why the interval instruments, which approximate the MPI, perform so well in our simulations. So far, the literature has exclusively exploited instruments that approximate the optimal instruments (Gandhi and Houde (2019), Reynaert and Verboven (2014)). We refer to these instruments as traditional instruments. These have been shown to work well in the usual Gaussian case. However, our simulations show that their performance declines when we depart from the Gaussian RC.

To evaluate the performance of our test and instruments, we conduct two sets of simulation experiments. First, we compare the performance of the test when using our interval instruments and when using the instruments commonly adopted by practitioners (Gandhi and Houde (2019), Reynaert and Verboven (2014)). We show that the test has the correct empirical size and that the interval instruments significantly outperform the traditional instruments in terms of power under alternative distributions. Second, we evaluate the performance of the interval instruments in estimating the model when the distribution of RC is flexibly parametrized, and follows a Gaussian mixture. We show that our instruments outperform the traditional instruments in terms of the mean squared error. In the case where the RC is a simple Gaussian, the three sets of instruments perform equally well.

Finally, we apply the tools developed in this paper to estimate the demand for cars in Germany from 2012 to 2018. The objective of the empirical exercise is to see how well our instruments perform at estimating a flexible distribution of RCs using a real dataset. Given the importance of price to address most empirical questions, we increase the flexibility of the model by estimating a Gaussian mixture for the RC associated with price. Second, we use our specification test to assess how the degree of misspecification decreases when we increase the flexibility in the distribution of RCs. Third, we use our results to study how the shape of

the RC on price can modify important counterfactual quantities such as the pass-through. In particular, our empirical results are consistent with the findings in Miravete et al. (2022).

Related literature Our paper contributes to several strands of the literature. First, it contributes to the literature on the flexible estimation of aggregate demand models for differentiated goods. A few recent papers have proposed non-parametric and semi-parametric methods to estimate aggregate demand functions. Compiani (2018) proposes a non-parametric estimator of the demand functions. If relaxing all the parametric assumptions makes this approach conceptually appealing, it also faces significant theoretical and practical difficulties (more stringent data requirements, large curse of dimensionality, limited scope for counterfactual analysis).[2] Lu, Shi, and Tao (2021) and Wang (2022) propose semi-parametric estimators of the distribution of RCs. These approaches are complementary to ours and the instruments we develop in this paper can be useful to implement their non-parametric IV estimation procedures, which are known to be rather sensitive to the quality of the instruments (Chetverikov and Wilhelm (2017)). Finally, Ho and Pakes (2014), Tebaldi, Torgovitsky, and Yang (2019) suggest deriving bounds directly on the counterfactual quantities.

Our paper also contributes to the literature on the non-parametric identification of the distribution of RCs in demand models (Fox and Gandhi (2011), Fox, il Kim, Ryan, and Bajari (2012), Dunker, Hoderlein, and Kaido (2022), Wang (2022), Berry and Haile (2014)). First, we slightly extend the identification result in Wang (2022) to link it directly to the primitives of the model, without assuming that the demand functions are identified. Second, we provide a practical way of constructing moments that feature high identifying power with respect to the distribution of RCs.

Third, we contribute to the literature that focuses on the practical estimation of the BLP model. First, we show that the interval instruments that we construct in this paper can be successfully used to estimate the distribution of random coefficients, and particularly so under of flexible distribution of RCs. This new set of instruments complements instruments commonly used by practitioners: Reynaert and Verboven (2014) and Gandhi and Houde (2019) (see Conlon and Gortmaker (2020) for a review). Moreover, we provide a new parametrization of the

[2] In particular, Compiani (2018) relaxes the Type 1 Extreme Value assumption on the taste shock. However, it is not clear how restrictive this assumption is. McFadden and Train (2000) shows that a mixed-logit model with flexibly distributed random coefficients can approximate any discrete choice model derived from random utility maximization. On the other hand, the Type 1 Extreme Value assumption generates massive computational gains, which allows for studying sophisticated markets with many products and many characteristics. Thus, the cost-benefit analysis seems to be largely in favor of the logit specification.

model, which facilitates the estimation when the distribution of RCs is a Gaussian mixture. This new parametrization complements previous papers that aim at improving the estimation of the model (Dubé, Fox, and Su (2012), Lee and Seo (2015), Salanié and Wolak (2019)).

Finally, our paper contributes to the literature on the asymptotic properties of the BLP estimator (Armstrong (2016), Berry, Linton, and Pakes (2004), Freyberger (2015), Ketz (2019)). In particular, we prove the asymptotic normality and the consistency of the BLP estimator in the large market framework under less stringent assumptions than the remainder of the literature.

Structure of the paper In Section 2, we recall the baseline BLP model, define the structural error of the model, and provide conditions under which the distribution of RCs is nonparametrically identified. In Section 3, we derive the most powerful instrument and show how it relates to the classical optimal instruments. In Section 4, we construct two feasible approximations of the MPI. In Section 5, we present our specification test and show its asymptotic validity. In Section 6, we conduct Monte Carlo simulations to evaluate the consequences of misspecification on quantities of interest, and gauge the performance of our test and instruments. In Section 7, we apply our new tools to estimate the demand for cars in Germany. We conclude the paper in section 8.

2 Model and identification

2.1 Indirect utility and moment restrictions

Indirect utility We first describe the indirect utility function that induces the observed market shares. Our setting closely follows the one introduced in the seminal paper Berry et al. (1995). There are T markets indexed by $t = 1, ..., T$. There is a continuum of consumers indexed by i. There are J_t market-specific products in market t. Each consumer chooses a product $j \in \{0, 1, ..., J_t\}$ where $j = 0$ corresponds to the outside option. For the sake of exposition and without loss of generality, we will assume throughout our analysis that the number of products is constant across markets ($\forall t,\ J_t = J$). Product j is characterized by a vector of characteristics x_{jt}, which includes the price of the good in most empirical settings. Consumer i derives an indirect utility u_{ijt} from purchasing good $j \in \{0, 1, \ldots, J\}$ in market t:

$$u_{ijt} = \underbrace{x'_{1jt}\beta + \xi_{jt}}_{\delta_{jt}} + x'_{2jt}v_i + \varepsilon_{ijt}, \tag{2.1}$$

with the following:

- x_{1jt} is a vector of product characteristics of dimension K_1 associated with product j and for which there is no preference heterogeneity; β represents preferences for x_{1jt};
- ξ_{jt} is an unobserved demand shock on product j in market t;
- $\delta_{jt} \equiv x'_{1jt}\beta + \xi_{jt}$ denotes the mean utility for product j, the part of the utility that is common to all consumers;
- x_{2jt} is a vector of product characteristics of dimension K_2 for which there is preference heterogeneity; v_i is the associated random coefficient that follows a distribution characterized by density f and is independent of all the other variables: $v_i \perp\!\!\!\perp (x_t, \xi_t, \{\varepsilon_{ijt}\}_{j=1,\dots J})$;
- ε_{ijt} is a preference shock that follows an Extreme Value type I (EV1) distribution independent of all other variables and across i, j, t.

For individual i in market t, the indirect utility from purchasing the outside option is normalized to $u_{i0t} = \varepsilon_{i0t}$. From the random utility functions in (2.1), we can infer the demand functions for each good j in market t denoted $\rho_{jt}(f, \beta)$. Each consumer chooses the product that maximizes his or her utility. Let y_{ijt} equal 1 if individual i chooses good $j = 0, 1, \dots, J$ in market $t = 1, \dots, T$. We have the following:

$$\begin{aligned}
\forall j \neq 0, \quad \rho_{jt}(f, \beta) &\equiv \mathbb{P}_{f,\beta}(y_{ijt} = 1 | x_t, \xi_t) \\
&= \mathbb{P}_{f,\beta}(\text{good } j \text{ is chosen in market } t \text{ by individual } i | x_t, \xi_t) \\
&= \mathbb{P}_{f,\beta}(u_{ijt} > u_{ikt} \ \forall k \neq j | x_t, \xi_t) \\
&= \int_{\mathbb{R}^{K_2}} \frac{\exp\left\{x'_{1jt}\beta + \xi_{jt} + x'_{2jt}v\right\}}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1kt}\beta + \xi_{kt} + x'_{2kt}v\right\}} f(v) dv.
\end{aligned} \tag{2.2}$$

For the outside option, the demand function is written as follows:

$$\rho_{0t}(f, \beta) = \mathbb{P}_{f,\beta}(y_{i0t} = 1 | x_t, \xi_t) = \int_{\mathbb{R}^{K_2}} \frac{1}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1kt}\beta + \xi_{kt} + x'_{2kt}v\right\}} f(v) dv.$$

Following the EV1 assumption on the idiosyncratic shock on utility, the demand functions take the usual logit form integrated over the distribution of preference heterogeneity. We assume in this paper that the observed market shares are equal to the shares generated by the model above at the true distribution f and the true preference parameter β:

$$\forall j, \ \ \forall t, \qquad s_{jt} = \rho_{jt}(f, \beta). \tag{2.3}$$

Moment restrictions Following the literature, we assume that the unobserved demand shock ξ_{jt} is mean independent of z_{jt}, a set of instrumental variables, namely, $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$ *a.s.*. The set z_{jt} traditionally consists of the exogenous characteristics of all the products on the market as well as cost shifters, which are meant to instrument for price. Indeed, the price of a good is usually considered to be an endogenous variable since it is correlated with the unobserved demand shock ξ_{jt} through the profit maximization problem of firms.[3] To estimate the model, the researcher chooses functions of the instruments z_{jt} to construct a set of unconditional moments. We refer to these functions as estimation instruments and denote them $h_E(z_{jt})$. Likewise, in our analysis, we study the functions of the instruments that are designed to test the specification of the model. We refer to these instruments as testing instruments and we denote them $h_D(z_{jt})$, where D stands for detection.

2.2 Inverse demand function and structural error

Inverse demand function For any given distribution of random coefficients $\tilde{f}$, we define the demand function $\rho \equiv (\rho_1(\cdot), ..., \rho_J(\cdot))$ as the function which maps the vector of mean utilities δ to the vector of market shares generated by the model under $\tilde{f}$:

$$\begin{aligned} \rho(\cdot, x_{2t}, \tilde{f}) : \ & \mathbb{R}^J \to [0,1]^J \\ & \delta \mapsto \int_{\mathbb{R}^{K_2}} \frac{\exp\left\{\delta + x_{2jt}' v\right\}}{1 + \sum_{k=1}^{J} \exp\left\{\delta_k + x_{2kt}' v\right\}} \tilde{f}(v) dv. \end{aligned}$$

[3] To deal with the endogeneity of prices, Berry et al. (1995) also suggests using exogenous own-product characteristics as well as exogenous characteristics from other products. The main idea behind the use of these instruments is to take advantage of the correlation between price and exogenous characteristics implied by profit-maximizing firms. To be precise, Berry et al. (1995) suggests using the sum of the characteristics from other products produced by the same firm and the sum of exogenous characteristics from rival firms' products as instruments.

Berry (1994) shows by applying Brouwer's fixed point that for any (s_t, x_{2t}) and for any distribution of random coefficients $\tilde{f}$ (even when $\tilde{f}$ is not the true distribution), there exists a unique $\tilde{\delta} \in \mathbb{R}^J$ such that:

$$s_t = \rho(\tilde{\delta}, x_{2t}, \tilde{f}).$$

We define the solution to the previous system of equations as the inverse demand functions: $\rho^{-1}(s_t, x_{2t}, \tilde{f}) = \tilde{\delta}$. Unfortunately, there is no closed form expression for the inverse demand function, which must be recovered numerically.

Structural error From what precedes, we can uniquely define the structural error $\xi_{jt}(\tilde{f}, \tilde{\beta})$ generated by a distribution of random coefficient $\tilde{f}$ and a homogeneous parameter $\tilde{\beta}$:

$$\xi_{jt}(\tilde{f}, \tilde{\beta}) = \rho_j^{-1}(s_t, x_{2t}, \tilde{f}) - x_{1jt}'\tilde{\beta}. \tag{2.4}$$

The non-linear nature of the model is captured by the inverse demand function which enters the expression of the structural error. The absence of an analytical formula for the inverse demand implies that there is no closed form expression for the structural error, which complicates the estimation of the BLP demand model. If we consider a parametric family of distributions $\tilde{\mathcal{F}} = \{\tilde{f}(\cdot|\tilde{\lambda}) : \tilde{\lambda} \in \tilde{\Lambda}\}$, then the structural error generated by a specific element in $\tilde{f}(\cdot|\tilde{\lambda}) \in \tilde{\mathcal{F}}$ and $\tilde{\beta}$ is defined as follows:

$$\xi_{jt}(\tilde{f}(\cdot|\tilde{\lambda}), \tilde{\beta}) = \rho_j^{-1}(s_t, x_{2t}, \tilde{f}(\cdot|\tilde{\lambda})) - x_{1jt}'\tilde{\beta}.$$

2.3 Non-parametric identification

The main objective of this paper is to provide tools to test the specification on the distribution of random coefficients and to improve its estimation under a flexible specification. A natural first step is to study the conditions under which this distribution is non-parametrically identified. The identification of random coefficients in multinomial choice models has been studied extensively in the literature (Allen and Rehbeck (2020), Berry and Haile (2014), Dunker et al. (2022), Fox and Gandhi (2011), Fox et al. (2012), Wang (2022)). We summarize some of these findings in Appendix C.1. In this Section, we build on an important identification result in Wang (2022) to recover a set of sufficient identifying conditions directly on the primitives of the model. We also show that the identification result holds with a less stringent exogeneity assumption than in Wang (2022).

In contrast to the rest of the literature, Wang (2022) adopts all the parametric assumptions in the standard BLP model and looks for a set of sufficient restrictions under which the

identification of the demand functions implies the identification of the distribution of random coefficients. This approach allows him to obtain conditions that are less stringent than the rest of the literature. In particular, Wang (2022) makes no special regressor assumption, no full support assumption, and no continuity assumption on the covariates. Specifically, he shows that if the demand functions $\rho = (\rho_1, ..., \rho_J)$ are identified on an open set of $\mathbb{R}^J$, then the distribution of random coefficients is identified.[4] His proof exploits the real analytic property of the demand functions.[5] In this paper, we build on this injectivity result to find sufficient identifying conditions directly on the primitives of the model (without assuming identification of the demand functions). We also show using a random permutation of the indices that we only require the demand shock ξ_{jt} to be mean independent of the instrumental variables z_{jt} across products, but we do not require this to hold for each product j taken separately. Formally, we only require $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$ *a.s.* and not $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$ *a.s.* for all product j as previously. This is less restrictive, as demand shocks can now be on average non-zero for certain products and account for unobserved quality inherent to each product.

Let us formally state the assumptions that we impose to recover the point identification of (f, β).

Assumption A

(i) Strict exogeneity: $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$ *a.s.;*

(ii) Completeness: for any measurable function g *such that* $\mathbb{E}[|g(s_t, x_t)|] < \infty$*, if* $\mathbb{E}[g(s_t, x_t)|z_{jt}] = 0$ *a.s., then* $g(s_t, x_t) = 0$ *a.s.;*

(iii) The distribution of the data $(s_t, x_{2t}, x_{1t}, z_t)$ *is fully observed by the econometrician and market shares* s_t *are generated by the demand model defined in section 2.1 by equations (2.1) and (2.3);*

(iv) Detectable difference in distributions: we say f *and* $\tilde{f}$ *differ (and write* $f \neq \tilde{f}$*) if there exists* $\bar{v} \in \mathbb{R}^{K_2}$ *such that* $F(\bar{v}) \neq \tilde{F}(\bar{v})$*;*

(v) Let $x_t = (x_{1t}, x_{2t})$ *then* x_t *is such that* $\mathbb{P}(x_t' x_t$ *is positive definite*$) > 0 \ \ \forall t$*;*

(vi) There exists $\bar{x}_t \in \mathcal{X}$ *and an open set* $\mathcal{D} \subset \mathbb{R}^J$ *such that* $\delta_t = \bar{x}_{1t}\beta_0 + \xi_t$ *varies on* $\mathcal{D}$ *a.s..*

In A(i), we assume that the instruments are strictly exogenous. Assumption A(ii) is a completeness assumption that states that the instruments are strongly relevant with respect to (s_t, x_t). This assumption is typical of semiparametric or nonparametric IV models and is

[4] Identification of demand functions can be achieved using Theorem 1 in Berry and Haile (2014).

[5] In particular, the real analytic property yields that the local identification of ρ on $\mathcal{D} \subset \mathbb{R}^J$ implies the identification of ρ on $\mathbb{R}^J$ From the global identification of ρ, he is then able to show that the random coefficients' distribution is identified under a simple rank condition on x_{2t}.

equivalent to a full rank assumption in a linear IV model. Intuitively, it means that if the inverse demands are different almost surely, then the instruments will be able to detect the difference. The completeness assumption is a strong assumption that has been widely used in this literature (Berry and Haile (2014), Dunker et al. (2022), Wang (2022)). Assumption A(v) is a standard rank condition. Assumption A(vi) is meant to ensure that there is enough variation in δ_t to apply the injectivity result in Wang (2022). This assumption indicates that there needs to be sufficient variation in product characteristics across markets in the data to identify f. In practice, product characteristics are very similar from one market to the other and may not yield sufficient variation. A judicious solution is to create inter-market variation by interacting product characteristics with demographic variables characterizing each market. Let us now state our formal identification result.

Proposition 2.1
Under Assumption A, the distribution of random coefficients f and the homogeneous preference parameters β are non-parametrically identified:

$$(\tilde{f},\tilde{\beta}) = (f,\beta) \iff \mathbb{E}[\xi_{jt}(\tilde{f},\tilde{\beta})|z_{jt}] = \mathbb{E}\left[\rho_j^{-1}(s_t,x_{2t},\tilde{f}) - x_{1jt}'\tilde{\beta}\middle| z_{jt}\right] = 0 \ \ a.s..$$

The proof is in Appendix B.1. The identification result above entails that under some fairly weak conditions and in the presence of instruments that generate sufficient variation in the product characteristics, the observed data identifies the distribution of random coefficients non-parametrically. Formally, the model is at the true pair (f,β) if and only if the associated structural error is mean independent of the instrumental variables z_{jt}. We use this identification result to show the consistency of our test under a specific choice of instruments that we will characterize thereafter.

3 Detecting misspecification: the most powerful instrument

The aim of this section is to recover the instrument with the greatest ability to detect misspecification in the distribution of RC. To do so, we consider a setting in which the econometrician wants to test a simple hypothesis of the form $\overline{H}_0 : (f,\beta) = (f_0,\beta_0)$. The upper bar is used to stress the fact that $\overline{H}_0$ is a simple hypothesis, in contrast to the composite hypothesis $H_0 : f \in \mathcal{F}_0$ that we study in section 5. Our approach builds on a simple intuition: if the model under $\overline{H}_0$ is misspecified, then the structural error will depart from the true demand shock ξ_{jt}, and our goal is to find the best instrument to pin down this deviation. We proceed as

follows. First, we introduce a moment-based test for $\overline{H}_0$ and we show its asymptotic validity. Next, we derive an analytical expression for the instrument that maximizes the power of our test against a fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$. We call this instrument the most powerful instrument (MPI) and we show how it relates to the classical optimal instruments, derived for efficient estimation purposes. In Section 4, we provide two feasible approximations of the MPI, which have the critical property of being invariant with respect to the alternative $\overline{H}_a$.

3.1 A moment-based test

We want to test $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$ against $H_a : (f, \beta) \neq (f_0, \beta_0)$. For any set of testing instruments $h_D(z_{jt})$, we have the following implication:

$$\overline{H}_0 : (f, \beta) = (f_0, \beta_0) \implies \overline{H}_0' : \quad \mathbb{E}[h_D(z_{jt})\xi_{jt}(f_0, \beta_0)] = 0.$$

We propose to test $\overline{H}_0$ indirectly through its implication $\overline{H}_0'$, which is a set of unconditional moment conditions. We test $\overline{H}_0'$ with a moment-based test. Our test statistic writes as follows:

$$S_T(h_D, f_0, \beta_0) = TJ \left(\frac{1}{TJ} \sum_{j,t} \xi_{jt}(f_0, \beta_0) h_D(z_{jt}) \right)' \hat{\Omega}_0^{-1} \left(\frac{1}{TJ} \sum_{j,t} \xi_{jt}(f_0, \beta_0) h_D(z_{jt}) \right), \tag{3.5}$$

with $\hat{\Omega}_0$ a consistent estimator of Ω_0 the asymptotic variance-covariance matrix of

$$\frac{1}{\sqrt{TJ}} \sum_{j,t} h_D(z_{jt}) \xi_{jt}(f_0, \beta_0)$$

which is $\Omega_0 = \mathbb{E}[\xi_{jt}^2(f_0, \beta_0) h_D(z_{jt}) h_D(z_{jt})']$. We study the asymptotic properties of our test as the number of markets, T, goes to infinity. As the focus of this section is on the construction of the most powerful instrument, we postpone the treatment of the specific challenges implied by parameter uncertainty (i.e. when β_0 and f_0 must be estimated beforehand) and by the numerical approximations involved in the derivation of the structural error (in practice, the researcher derives a numerical approximation of $\xi_{jt}(f_0, \beta_0)$) to Section 5. Additionally, to keep the results as simple as possible while retaining the key intuitions, we assume independence of the demand shocks in a given market conditional on z_{jt}. This last assumption is relaxed in the proofs in Appendix B.2 and in section 5.

Proposition 3.1
Assume that (s_t, x_t, z_t) *are i.i.d. across markets and consistent with the probability model defined by equations (2.1), (2.2) and (2.3) evaluated at* (f, β), $\mathbb{E}[\|\xi_{jt}(f_0, \beta_0) h_D(z_{jt})\|^2] < +\infty$, Ω_0 *has full rank, and, for* $k \neq j$, $\xi_{jt} \perp\!\!\!\perp \xi_{kt} | z_t$. *We have the following:*

- *under* $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$, $\quad S_T(h_D, f_0, \beta_0) \xrightarrow[T \to +\infty]{d} \chi^2_{|h_D|_0}$,
- *under* $H'_a :$ $\mathbb{E}\left[h_D(z_{jt}) \xi_{jt}(f_0, \beta_0)\right] \neq 0$, $\quad \forall q \in \mathbb{R}^+$, $\mathbb{P}(S_T(h_D, f_0, \beta_0) > q) \underset{T \to +\infty}{\longrightarrow} 1$,

with $|\cdot|_0$ *being the counting norm.*

The previous proposition indicates that as long as the testing instruments are functions of z_{jt}, our test procedure is asymptotically valid for $\overline{H}_0$. We are testing $\overline{H}_0$ by virtue of its implication $\overline{H}'_0 : \mathbb{E}\left[h_D(z_{jt}) \xi_{jt}(f_0, \beta_0)\right] = 0$ and, as a consequence, the power properties of our test hinge critically on the choice of the testing instruments $h_D(z_{jt})$. This is the focus of the next subsection.

3.2 The most powerful instrument (MPI)

The choice of testing instruments $h_D(z_{jt})$ is key to maximizing the rejection rate of $\overline{H}_0$ under any alternative $H_a : (f, \beta) \neq (f_0, \beta_0)$. To guide our choice of instruments, we first derive the instrument that maximizes the power of our test when the econometrician tests $\overline{H}_0$ against a fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a) \neq (f_0, \beta_0)$. We refer to this instrument as the most powerful instrument (MPI). In practice, the researcher is often reluctant to fix the alternative. However, the MPI represents a useful first-best solution for which we provide two feasible approximations in section 4.

Derivation of the most powerful instrument To construct the MPI, we use the following decomposition of the structural error generated under $\overline{H}_a$:

$$\xi_{jt}(f_0, \beta_0) = \underbrace{\xi_{jt}(f_a, \beta_a)}_{\text{true error under } \overline{H}_a} + \underbrace{\xi_{jt}(f_0, \beta_0) - \xi_{jt}(f_a, \beta_a)}_{\Delta^{\xi_{jt}}_{0,a}},$$

with $\Delta^{\xi_{jt}}_{0,a}$ being the correction term due to misspecification under the alternative $\overline{H}_a$. Our goal is to compare the ability of our test for different candidates $h_D(z_{jt})$, to reject $\overline{H}_0$ under $\overline{H}_a$.

The literature offers many ways to compare the power of competing tests (see Gourieroux and Monfort (1995) for a comprehensive review). First, we distinguish between exact and approximate methods. Exact methods rely on the exact distribution of the test statistic (under $\overline{H}_0$) and allow for comparison in finite sample while asymptotic methods exploit the asymptotic distribution of the test statistic and are informative in larger samples. In our case, the exact distribution of our test is unknown. Thus, we rely on asymptotic methods, which is the most common case in the literature. Second, we divide the methods into local and non-local methods. In parametric tests, local strategies are based on the analysis of the power properties of competing tests under a sequence of local alternatives θ_T which converges to θ_0 at a given rate (usually $\frac{1}{\sqrt{T}}$). The econometrician can compare two competing tests by means of their power functions (or more precisely, the limits of these power functions when sample sizes go to $+\infty$). This is called the direct approach. The dual approach, which is known as Pitman's relative efficiency, consists of comparing the rates at which the minimal number of observations must increase to ensure a given level of power. The approach we favor in this paper is the non-local approach developed in Bahadur (1960). Here, the econometrician chooses the test with the smallest level α needed to attain a given power against a fixed alternative and for a given number of observations. In other words, the econometrician chooses the test that minimizes the risk of type I error ceteris paribus.

There are several reasons to favor Bahadur's non-local approach. First, it is better suited for the testing problem we study in this paper. The comparison criterion, known as the asymptotic slope of the test, is in our case straightforward to derive, whereas it is not clear how one should derive Pitman's efficiency criterion when the test concerns non-parametric objects such as distributions. Moreover, we study the properties of our test against a fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$ as in Bahadur's case, which is not necessarily local. Finally, the literature has highlighted many limitations of the local approach. Local criteria are often unable to discriminate between tests even when these tests lead to different decisions (see Silvey (1959)). In addition, as shown in Dufour and King (1991), a locally optimal test in a neighborhood of H_0 may perform very poorly away from H_0.

Let us now present the intuition for Bahadur's comparison approach. From Section 3.1, we have:

$$\text{Under } \overline{H}_0: \qquad S_T \equiv S_T(h_D, f_0, \beta_0) \xrightarrow{d} S \qquad \text{with } \; S = \chi^2_{|h_D|_0}.$$

Following the same notations as in Gourieroux and Monfort (1995), we denote:

$$\Lambda(s) = \mathbb{P}_{\overline{H}_0}(S \geq s).$$

The critical value is usually derived using the asymptotic distribution of the test statistic under H_0. The approximate critical region at a given level α is then given by:

$$CR_\alpha = \{S_T \geq \Lambda^{-1}(\alpha)\} = \{\Lambda(S_T) \leq \alpha\}.$$

The main idea in Bahadur's approach entails deriving the level of the test if one takes the value of the test statistic as the critical value (this is also known as the p-value). Namely:

$$\alpha_T = \Lambda(S_T).$$

Bahadur suggests preferring the test that displays the lowest level α_T at least asymptotically. A formal analysis of the asymptotic behavior of α_T shows that it is better to consider the limit of a transformation of α_T than the limit of α_T directly. This gives rise to the concept of the approximate slope of the test.

Definition 1 (Asymptotic slope of the test)

(i) $K_T = -\frac{2}{T}\log(\Lambda(S_T))$ is the approximate slope of the test,

(ii) Under $\overline{H}_a$: $plim\ K_T = c(f_a, \beta_a)$ is the asymptotic slope of the test,

with $plim$, the limit in probability when $T \to +\infty$.

Under the alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$, consider two sequences of tests based on S_T^1 and S_T^2 with asymptotic slopes $c^1(f_a, \beta_a)$ and $c^2(f_a, \beta_a)$ respectively. The test based on S_T^1 is asymptotically preferred to the test based on S_T^2 in Bahadur's sense if and only if $c^1(f_a, \beta_a) > c^2(f_a, \beta_a)$. To derive the asymptotic slopes of our test, we apply an important result in Geweke (1981), which states that if under H_0: $S_T \xrightarrow[T\to+\infty]{d} \chi_q^2$ (with any $q \in \mathbb{N}^*$), then $\frac{1}{T}S_T \xrightarrow{a.s.} c(f_a, \beta_a)$ (when the limit exists). In our test, the limiting distribution is chi-squared. Thus, the asymptotic slope of our test with instrument $h_D(z_{jt})$ writes:

$$c_{h_D}(f_a, \beta_a) = plim\ \frac{1}{T}S_T(h_D, f_0, \beta_0) = J\mathbb{E}\left[\xi_{jt}(f_0, \beta_0)h_D(z_{jt})\right]' \Omega_0^{-1}\mathbb{E}\left[\xi_{jt}(f_0, \beta_0)h_D(z_{jt})\right].$$

Let us note that the asymptotic slope can also be interpreted as a measure of the speed of divergence of the test statistic in terms of population moments, i.e. speed of divergence $\approx T \times c_{h_D}(f_a, \beta_a)$. In the next proposition, we derive an analytical expression for the instrument that maximizes the slope of the test.

Proposition 3.2 (Most powerful instrument)
Let $\mathcal{H}$ be the set of measurable vectorial functions of z_{jt}. Under any fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$, we have the following:

$$\left(\mathbb{E}\left[\xi_{jt}(f_0, \beta_0)^2 | z_t\right]\right)^{-1} \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}] \in \underset{h_D \in \mathcal{H}}{argmax} \; c_{h_D}(f_a, \beta_a).$$

The proof is given in Appendix B.2. The MPI equals the conditional expectation of the correction term $\Delta_{0,a}^{\xi_{jt}}$ divided by a conditional variance term $\mathbb{E}\left[\xi_{jt}(f_0, \beta_0)^2 | z_{jt}\right]$. For exposition purposes, we drop the conditional variance term in the subsequent analysis and take the homoskedastic MPI $h_D^*(z_{jt}) = \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]$ as the reference MPI.[6] Methods have been proposed to estimate the conditional variance term non-parametrically and could be adapted to our case. However, it is well known that conditional variance, which also appears in the formulation of the optimal instruments, is difficult to model and estimate in practice. In the BLP framework, the large dimension of z_{jt} makes the exercise even more difficult. Hence, researchers typically ignore this term or impose a restrictive and ad-hoc structure on the form that it can take (for instance, Reynaert and Verboven (2014)'s approximation of the optimal instruments in the BLP model ignores the variance term). The homoskedastic MPI, $h_D^*(z_{jt})$, features other appealing properties including (i) consistency of the associated test and (ii) maximizing correlation with the structural error under the alternative.[7] For simplicity, in what follows, we refer to the homoskedastic MPI as the MPI.

(i) Consistency By setting h_D equal to h_D^*, our moment-based test becomes consistent against any fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a) \neq (f_0, \beta_0)$. Namely, we have the following result:

Proposition 3.3 (Consistency of the test with the MPI)
Under Assumption A and the same assumptions as in Proposition 3.1, we have:

$$\overline{H}_a : (f, \beta) = (f_a, \beta_a) \neq (f_0, \beta_0) \implies \forall q \in \mathbb{R}^+, \; \mathbb{P}(S_T(h_D^*, f_0, \beta_0) > q) \underset{T \to +\infty}{\longrightarrow} 1.$$

The proof of this result is given in Appendix B.2.

[6] This last expression corresponds to the exact formulation of the MPI under homoskedasticity.
[7] The consistency of the test also holds when we keep the conditional variance term.

(ii) Correlation with the structural error Another interesting property of the MPI is to be the function of z_{jt} which maximizes the correlation with the structural error.

Proposition 3.4 (Correlation between the MPI and the structural error)
Let $\mathcal{H}$ be the set of measurable functions of z_{jt}, we have under $\overline{H}_a$:

$$\forall \alpha \in \mathbb{R}^*, \quad \alpha \, \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}] \in \underset{h \in \mathcal{H}}{\arg\max} \; \left| corr(\xi_{jt}(f_0, \beta_0), h(z_{jt})) \right|.$$

The proof is given in Appendix B.2. Intuitively, the MPI $h_D^*(z_{jt})$ is designed to fully capture the exogenous variation contained in the correction term $\Delta_{0,a}^{\xi_{jt}}$ implied by the misspecification, which yields the result above.

3.3 Connection with the optimal instruments

The MPI maximizes the power of the moment-based test for $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$. In contrast, the optimal instruments minimize the asymptotic variance-covariance of the GMM estimator when the parameter of interest is identified by conditional moment restrictions. These two problems are seemingly unrelated. However, we show that the MPI devoted to testing the specification of the model at the true parameter against any fixed local alternative can be rewritten as a linear combination of the optimal instruments. Consequently, one can reinterpret the optimal instruments as a local approximation of the MPI devoted to testing the model at the true parameter. This connection between the MPI and the optimal instruments helps us understand why the feasible approximations of the MPI we construct in section 4 improve the performance of the BLP estimator in our Monte Carlo simulations when the distribution of RCs is flexible. In this subsection, we first derive the optimal instruments. Then, we exhibit the relation between the optimal instruments and the MPI.

The estimation of the model works as follows. The researcher assumes that f belongs to a parametric family $\mathcal{F}_0 = \{f_0(\cdot|\tilde{\lambda}) : \tilde{\lambda} \in \Lambda_0\}$ and wants to estimate the true parameter $\theta_0 = (\beta_0', \lambda_0')'$ under this parametric restriction. In the estimation context that we study here, θ_0 refers to the true parameter. For now, let us assume that the model is correctly specified: $f \in \mathcal{F}_0$ and we shorten the notations by removing the dependence of the structural error in $f_0(\cdot|\tilde{\lambda})$, which becomes implicit in this context. Namely, $\xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta})$ becomes $\xi_{jt}(\tilde{\theta})$. We further

assume that θ_0 is point identified by the following moment restriction: $\mathbb{E}[\xi_{jt}(\theta_0)|z_{jt}] = 0$ *a.s.*.[8] The researcher must choose the set of instruments $h_E(z_{jt})$ (or equivalently, the unconditional moments) to include in the GMM objective function:

$$\hat{\theta} = \underset{\tilde{\theta}}{\text{Argmin}}\ TJ \left(\frac{1}{TJ}\sum_{j,t}\hat{\xi}_{jt}(\tilde{\theta})h_E(z_{jt})\right)' \hat{W} \left(\frac{1}{TJ}\sum_{j,t}\hat{\xi}_{jt}(\tilde{\theta})h_E(z_{jt})\right).$$

Optimal instruments in the BLP demand model Traditionally, the instruments $h_E(z_{jt})$ are chosen to minimize the asymptotic variance-covariance of the estimator $\hat{\theta}$. The instruments that reach this objective are called the optimal instruments. The resulting estimator is said to be efficient in the sense that its asymptotic variance cannot be reduced by using additional moment conditions. There is a large body of literature on the derivation of optimal instruments in econometric models (Amemiya (1974), Chamberlain (1987), Newey (1990, 2004)). The BLP estimator $\hat{\theta}$ is a non-linear GMM estimator and classical results in Chamberlain (1987) and Amemiya (1974) show that the optimal instruments in this case write:

$$h_E^*(z_{jt}) = \mathbb{E}[\xi_{jt}(\theta_0)^2|z_{jt}]^{-1}\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta_0)}{\partial \tilde{\theta}}\middle| z_{jt}\right],$$

The corresponding efficiency bound (obtained by setting $h_E = h_E^*$) writes:

$$V^* = \mathbb{E}\left[\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta_0)}{\partial \tilde{\theta}}\middle| z_{jt}\right]\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta_0)}{\partial \tilde{\theta}}\middle| z_{jt}\right]'\mathbb{E}[\xi_{jt}(\theta_0)^2|z_{jt}]^{-1}\right]^{-1}.$$

For the sake of exhaustivity, we show this result in Appendix B.2.1. As for the MPI, the formulation of the optimal instruments above is obtained under the assumption of conditional independence of demand shocks ξ_{jt} in the same market: $k \neq j$, $\xi_{jt} \perp\!\!\!\perp \xi_{kt}|z_t$. In Appendix B.2.1, we derive the expression for the optimal instruments under weaker assumptions on the demand shock.[9] Consistent with what we did in the case of the MPI, we drop the conditional variance term $\mathbb{E}[\xi_{jt}(\theta_0)^2|z_{jt}]$.

[8] The identification conditions in the parametric case are less stringent than the conditions for the non-parametric identification in Assumption A.

[9] We allow for unrestricted forms of correlation between demand shocks within a given market.

Connection between the MPI and the optimal instruments Let θ_0 the true parameter. Under the parametric assumption $f \in \mathcal{F}_0$, the simple hypothesis $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$ we studied previously becomes $\overline{H}_0 : \theta = \theta_0$. It is straightforward to show that, in the parametric case, the associated MPI against a fixed alternative $\overline{H}_a : \theta = \theta_a$ writes: $h_D^*(z_{jt}) = \mathbb{E}\left[\Delta_{\theta_0,\theta_a}^{\xi_{jt}} | z_{jt}\right]$ with $\Delta_{\theta_0,\theta_a}^{\xi_{jt}} = \xi_{jt}(\theta_0) - \xi_{jt}(\theta_a)$. By taking a Taylor expansion of $\xi_{jt}(\theta_a)$ around θ_0, we obtain the following:

$$\Delta_{\theta_0,\theta_a}^{\xi_{jt}} = \frac{\partial \xi_{jt}(\theta_0)}{\partial \tilde{\theta}}(\theta_0 - \theta_a) + o(||\theta_0 - \theta_a||_2)\,.$$

We see that when θ_a is in a neighborhood of θ_0, the MPI, $h_D^*(z_{jt})$, against this fixed alternative is a linear combination of the optimal instruments $h_E^*(z_{jt})$:

$$h_D^*(z_{jt}) = \mathbb{E}\left[\Delta_{\theta_0,\theta_a}^{\xi_{jt}} | z_{jt}\right] \approx \underbrace{\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta_0)}{\partial \tilde{\theta}}\middle| z_{jt}\right]'}_{h_E^*(z_{jt})}(\theta_0 - \theta_a).$$

It follows that classical optimal instruments can be interpreted as an approximation of the MPI devoted to testing $H_0 : \theta = \theta_0$ against any fixed local alternative.[10] Moreover, let us note that the connection between the MPI and the optimal instruments holds if we keep the conditional variance term in both cases.

4 A feasible most powerful instrument

The MPI is the most powerful instrument to reject $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$ against a fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$. Its derivation requires the knowledge of the alternative while in practice the econometrician typically wants to remain agnostic about the alternative. Moreover, the MPI is defined as a conditional expectation of a non-linear function with respect to a large dimension vector z_{jt}, and thus, even if the alternative $\overline{H}_a$ is known, the MPI can be difficult to compute. In this section, we remain in the same configuration, where the econometrician wants to test $\overline{H}_0 : (f, \beta) = (f_0, \beta_0)$ against a fixed alternative $\overline{H}_a : (f, \beta) = (f_a, \beta_a)$. However now, we assume that this alternative is unknown to the econometrician. We provide two feasible approximations of the MPI, which do not depend on $\overline{H}_a$, and that, unlike

[10] This interpretation of the optimal instruments only holds when the model is well specified i.e. $f \in \mathcal{F}_0$, and thus, in general, the optimal instruments shouldn't be used to test the specification of the model.

the MPI, can be computed in practice. To do so, we show that the MPI can be approximated by a linear combination of known functions of z_{jt}. We call these interval instruments in reference to the way these functions are derived. Our feasible MPI is simply the vector of the interval instruments. The cost to incur for feasibility is that the properties we established for the MPI do not carry over to the feasible MPI. Nevertheless, our Monte Carlo simulations in section 6 show that the interval instruments perform very well in practice.

By construction, in the BLP demand model, the correction term writes:

$$\begin{aligned}\Delta_{0,a}^{\xi_{jt}} &= x_{1jt}'(\beta_a - \beta_0) + \rho_j^{-1}(s_t, x_{2t}, f_0) - \rho_j^{-1}(s_t, x_{2t}, f_a) \\ &= x_{1jt}'(\beta_a - \beta_0) + \Delta_j(s_t, x_{2t}, f_0, f_a).\end{aligned} \tag{4.6}$$

The previous equation shows that the correction term is the sum of a linear part, which is standard, and a non-linear part which is specific to the BLP demand model.

Linear part The linear part of the MPI writes: $\mathbb{E}[x_{1jt}|z_{jt}]'(\beta_a - \beta_0) = \mathbb{E}[x_{1jt}|z_{jt}]'\gamma$. Thus, for its linear part, the MPI is a linear combination of the conditional expectation of x_{1jt} with respect to the exogenous variables with unknown weights. If one is interested in specifically testing that $\beta = \beta_0$, informative instruments simply consist of the variables in $\mathbb{E}[x_{1jt}|z_{jt}]$.

Non-linear part The non-linear part, $\Delta_j(s_t, x_{2t}, f_0, f_a)$, is the part which is implied by the misspecification on the distribution of RCs and for which we need to recover a feasible approximation. Equation (4.6) indicates that the non-linear part is the difference between the inverse demand functions generated by f_0 and f_a. We now go one step further and derive two analytical approximations of $\Delta_j(s_t, x_{2t}, f_0, f_a)$ which we then use as building blocks to construct our feasible approximations of the MPIs. The first approximation is based on a local expansion around f_0. The second approximation is based on an identity that is valid everywhere. The first approximation is more precise locally whereas the second one is more robust to large deviations from f_0.

4.1 Local approximation

First, we consider a local approximation of $\Delta_j(s_t, x_{2t}, f_0, f_a)$. This approximation corresponds to the first order term in the expansion of $\Delta(s_t, x_{2t}, f_0, f_a)$ "around f_0", which is recovered by exploiting the properties of the inverse demand function, which is both $\mathcal{C}^\infty$ and bijective in s_t.

Proposition 4.1
A first order expansion of $\Delta(s_t, x_{2t}, f_0, f_a)$ *around* f_0 *writes:*

$$\Delta(s_t, x_{2t}, f_0, f_a) = \left(\frac{\partial \rho(\delta_t^0, x_{2t}, f_0)}{\partial \delta}\right)^{-1} \int_{\mathbb{R}^{K_2}} \left[\frac{\exp(\delta_t^0 + x_{2t}v)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt}^0 + x_{2kt}'v\right\}} - \rho(\delta_t^0, x_{2t}, f_0)\right] f_a(v) + \mathcal{R}_0,$$

with $\delta_t^0 = \rho^{-1}(s_t, x_{2t}, f_0)$ *and* $\mathcal{R}_0 = o\big(\int_{\mathbb{R}^{K_2}} |f_a(v) - f_0(v)| dv\big)$.

The proof is in Appendix B.3.1. We first observe that for any density f_0, we can construct artificial market shares s_t^0 such that $\rho^{-1}(s_t, x_{2t}, f_a) = \rho^{-1}(s_t^0, x_{2t}, f_0)$. Then, we recover the final result by taking a Taylor expansion of $\rho^{-1}(s_t^0, x_{2t}, f_0)$ around s_t and showing that the remainder is bounded.[11] This approximation is local by design: it works best when f_a is a local deviation from f_0, even if it can be used more generally. To make this expression useful in practice, we must still overcome two difficulties. The distribution f_a is unknown to the econometrician. In addition, some variables such as δ_{jt}^0 are endogenous. However, notice that the previous expression may be particularly useful if the econometrician is interested in testing $\overline{H}_0$ against a fixed and known alternative as we did in the previous section.

Discretizing the integral To solve for the fact f_a is unknown to the econometrician, we replace the integral in which f_a appears by a finite Riemann approximation. Namely,

$$\int_{\mathbb{R}} \frac{\exp\left\{x_{2jt}'v\right\}}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt}^0 + x_{2kt}'v\right\}} f_a(v) dv \approx \sum_{l=1}^{L} \omega_l(f_a) \frac{\exp(x_{2jt}'v_l)}{1 + \sum_{k=1}^{J} \exp(\delta_{kt}^0 + x_{2kt}'v_l)},$$

with $\{v_l\}_{l=1,\dots,L}$ the points chosen in the domain of definition of f_a, and $\{\omega_l(f_a)\}_{l=1,\dots,L}$ the associated weights.[12] We provide more details on how to choose the points in Appendix C.4. It is important to observe that in the Riemann approximation, only the weights depend on the alternative f_a. This approximation can also be interpreted as approaching a continuous distribution with a discrete one, where each point in $\{v_l\}_{l=1,\dots,L}$ represents a specific consumer type with an associated probability $w_l(f_a)$. The non-linear part of the MPI can thus be approximated as follows:

[11] The expansion is taken around s_t because s_t^0 depends on f_a and is thus unknown to the researcher.

[12] In the usual Riemann sum, the weights correspond to density evaluated at point $v_l : f_a(v_l)$ times the width of the interval around v_l.

$$\mathbb{E}[\Delta_j(s_t, x_{2t}, f_0, f_a)|z_{jt}] \approx \sum_{l=1}^{L} \omega_l(f_a)\ \mathbb{E}[\pi_{j,l}(s_t, x_t)|z_{jt}],$$

$$\text{with } \pi_{j,l}(s_t, x_t) = \left(\frac{\partial \rho(\delta_t^0, x_{2t}, f_0)}{\partial \delta}\right)^{-1} \left[\frac{\exp(\delta_t^0 + x_{2t} v_l)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt}^0 + x_{2kt}' v_l\right\}} - \rho(\delta_t^0, x_{2t}, f_0)\right]_j.$$

Approximating the conditional expectation Ideally, we would like to estimate the conditional expectation of $\pi_{j,l}(s_t, x_t)$ with respect to z_{jt}. The endogenous variables are $\{\delta_{jt}^0\}_{j=1,....,J}$, and the potential endogenous variables in $\{x_{2jt}\}_{j=1,...,J}$, which often include prices. In practice, computing the conditional expectation is challenging because the dimension of z_{jt} can be very large and the functions $\pi_{j,l}(\cdot)$ are highly non-linear and non-separable in the endogenous variables. This makes it unappealing to use standard non-parametric estimation methods.[13] In the same spirit as Reynaert and Verboven (2014), we first project the endogenous variables on the space spanned by a relevant subset of z_{jt}. We mark the projected endogenous variables with a hat and we plug them into our functions $\pi_{j,l}(\cdot)$. Namely, we have the following approximation for every interval instrument l:

$$\mathbb{E}[\pi_{j,l}(s_t, x_t)|z_{jt}] \approx \hat{\pi}_{j,l}(z_{jt}) = \left(\frac{\partial \rho(\hat{\delta}_t^0, \hat{x}_{2t}, f_0)}{\partial \delta}\right)^{-1} \left[\frac{\exp(\hat{\delta}_t^0 + \hat{x}_{2t} v_l)}{1 + \sum_{k=1}^{J} \exp\left\{\hat{\delta}_{kt}^0 + \hat{x}_{2kt}' v_l\right\}} - \rho(\hat{\delta}_t^0, \hat{x}_{2t}, f_0)\right]_j.$$

We show in Appendix C.2 that this strategy yields an estimator of the conditional expectation that converges faster to a first order approximation of the conditional expectation.

Test procedure From what precedes, the MPI (for its non-linear part) can be approximated as follows: $h_D^*(z_{jt}) \approx \sum_{l=1}^{L} \omega_l(f_a)\ \hat{\pi}_{j,l}(z_{jt})$. As we don't know the weights $\omega_l(f_a)$, we propose to take the vector $\hat{\pi}_j(z_{jt}) = (\hat{\pi}_{j,1}(z_{jt}), ..., \hat{\pi}_{j,L}(z_{jt}))'$ as our testing instruments. We call them interval instruments in reference to the way we divide the support into several intervals to construct this approximation. Following the test procedure presented in section 3.1, we perform a moment based test for $\overline{H}_0 : \mathbb{E}\left[\hat{\pi}_j(z_{jt})\xi_{jt}(f_0, \beta_0)\right] = 0$. Under the same assumptions as in Proposition 3.1 and setting $h_D(z_{jt}) = \hat{\pi}_j(z_{jt})$, we have the following:

[13] For instance, a Sieve nonparametric estimator of the conditional mean. The dimension of z_{jt} makes this approach of little relevance in practice.

$$\text{Under } H_0 : S_T(h_D, f_0, \beta_0) \xrightarrow[T\to+\infty]{d} \chi^2_L.$$

This approach has the advantage of being feasible since we can construct the vector of interval instruments $\hat{\pi}_j(z_{jt})$, while remaining completely agnostic about f_a. The price to pay is that we lose the optimality properties of the MPI. We further discuss the properties of the feasible MPI in Appendix C.7. Moreover, the infeasible MPI, $h^*_D(z_{jt})$, is of dimension one and its test statistic is distributed as χ^2_L asymptotically. In contrast, the feasible MPI is of dimension L and its asymptotic distribution is a χ^2_L. This increase in the number of degrees of freedom may lead to some loss of power. An alternative approach would consist in letting the researcher choose the weights $\{\hat{\omega}_l\}_{l=1,\dots,L}$ and recover an instrument of dimension one. However, for this approach to work well and retain good power properties, the econometrician must choose the weights so that they approximately match the real weights $\{w_l(f_a)\}_{l=1,\dots,L}$. This requires a good prior knowledge of the cumulative distribution function of the alternative distribution f_a. Nevertheless, our Monte Carlo simulations in section 6 show that the feasible MPIs that we propose perform very well in practice.

4.2 Global approximation

Second, we consider a global approximation that is based on an identity which is valid everywhere and not only when f is close to f_a. Simple algebraic operations (see Appendix B.3.2) allow us to derive the following expression for $\Delta_j(s_t, x_{2t}, f_0, f_a)$. Let $\delta^0_{jt} = \rho_j^{-1}(s_t, x_{2t}, f_0)$ and $\delta^a_{jt} = \rho_j^{-1}(s_t, x_{2t}, f_a)$. We have:

$$\Delta_j(s_t, x_{2t}, f_0, f_a) = \log\left(\frac{\int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{2jt}v)}{1+\sum_{k=1}^J \exp\left\{\delta^a_{kt}+x'_{2kt}v\right\}} f_a(v)dv}{\int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{2jt}v)}{1+\sum_{k=1}^J \exp\left\{\delta^0_{jt}+x'_{2kt}v\right\}} f_0(v)dv}\right).$$

As for the local approximation, we cannot directly exploit this formula as some quantities such as f_a and δ^a_{jt} are unknown and some variables such as δ^0_{jt} are endogenous. To remedy these two difficulties, we apply the same methods as previously described: we discretize the integral, and we project the endogenous variables onto the space spanned by a relevant subset of z_{jt}. To solve for the fact that the mean utility δ^a_{jt} under the alternative is unknown, we replace it with the mean utility under the null δ^0_{jt}. This should not alter the approximation too much given that δ^a_{jt} only enters the expression at the denominator within a sum, which

averages out the differences between δ_{jt}^a and δ_{jt}^0 across products. In the end, we are able to provide the following approximation for the non-linear part of the MPI:

$$\mathbb{E}[\Delta_j(s_t, x_{2t}, f_0, f_a)|z_{jt}] \approx \log\left(\sum_{l=1}^{L} \bar{\omega}_l(f_a)\ \hat{\bar{\pi}}_{j,l}(z_{jt})\right) \text{ with } \hat{\bar{\pi}}_{j,l}(z_{jt}) = \frac{\frac{\exp(x'_{2jt}v_l)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\delta}_{kt}^0 + x'_{2kt}v_l\right\}}}{\int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{2jt}v)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\delta}_{jt}^0 + x'_{2kt}v\right\}} f_0(v)dv},$$

where $\{\bar{\omega}_l(f_a)\}_{l=1,...,L}$ correspond to the unknown weights and the $\hat{\bar{\pi}}_{j,l}(z_{jt})$ are set of global interval instruments. The MPI can thus be approximated by the logarithm of a weighted sum of known functions of z_{jt}. As we did previously, we use $\hat{\bar{\pi}}_j(z_{jt}) = (\hat{\bar{\pi}}_{j,1}(z_{jt}), ..., \hat{\bar{\pi}}_{j,L}(z_{jt}))'$ as instruments to test $\bar{H}_0$. All the weights are positive and sum to one, which entails that the non-linear part of the correction term is an increasing function of our instruments. This approximation is said to be global because contrary to the second approximation we study, it does not require f_0 to be close to f_a. Nevertheless, if f_a is close to f_0, then the fraction κ inside the logarithm is close to 1 and the well-known approximation $\log(\kappa) \approx \kappa - 1$ allows us to directly rewrite the MPI as a linear combination of our instruments.

Overall, the feasible MPIs that we derive in this section allows us to approximate the most powerful instrument against a fixed alternative while remaining agnostic about this alternative.

4.3 Feasible MPIs for estimation

In the estimation framework, the researcher stipulates that f belongs to a parametric family $\mathcal{F}_0 = \{f_0(\cdot|\tilde{\lambda}) : \tilde{\lambda} \in \Lambda_0\}$ and wants to estimate the true parameter $\theta_0 = (\beta_0', \lambda_0')'$ under this parametric restriction. From the connection between the MPI and the local instruments that we present in section 3.3, we can infer that good estimation instruments $h_E(z_{jt})$ ought to approximate the MPI devoted to testing $H_0 : \theta = \theta_0$ against any local alternative. If we have an initial estimator of θ_0, we can directly use the interval instruments presented previously to approximate the MPI devoted to testing $H_0 : \theta = \theta_0$ against an unknown alternative. The fact that the feasible MPIs do not depend on the alternative is key for estimation. Moreover, the transformation of the MPI into a vector of instruments of dimension $L \geq |\lambda_0|$ is necessary for estimation as the number of instruments must be greater than the dimension of the parameter to estimate.[14] In Appendix C.5, we propose a version of the interval instruments that does not

[14]The linear parameter β_0 has its own instruments, which are simply the variables in x_{1jt}.

require a first step estimate of θ_0 and that can be computed directly from the logit specification.

5 Composite hypothesis

In the traditional estimation procedure, which encompasses almost all the applications of the BLP model, the econometrician must make a parametric assumption on the distribution of random coefficients to estimate the model. Formally, the econometrician assumes f belongs to a parametric family $\mathcal{F}_0 = \{f_0(\cdot|\tilde{\lambda}) : \tilde{\lambda} \in \Lambda_0\}$, where $\tilde{\lambda}$ is a parameter that must be estimated. In applied work, researchers typically assume that f is normally distributed. This parametric choice is rarely grounded in economic theory and, if too restrictive, is likely to impose arbitrary restrictions on some key counterfactual quantities such as the pass-through. In this section, we develop a formal specification test for $H_0 : f \in \mathcal{F}_0$. In comparison to the test in section 3.1, we must now estimate the parameters of the distribution $\theta_0 = (\beta_0', \lambda_0')'$ in a first step, which generates parameter uncertainty. Moreover, we propose a rigorous treatment of the numerical approximations involved in the derivation of the structural error $\xi_{jt}(\tilde{\theta})$. We organize this section as follows. First, we define the pseudo-true value associated with a given specification and the first stage estimator. Second, we define our test procedure and its implementation in practice. Finally, we study the asymptotic properties of our test.

5.1 Pseudo-true value and first stage estimator

To estimate the BLP model, researchers must make three choices. They must choose the parametric family $\mathcal{F}_0$, the instruments $h_E(z_{jt})$ to estimate the model, and a weighting matrix W, which weights the different moments included in the objective function. Given these three choices, we can define the BLP pseudo-true value $\theta(\mathcal{F}_0, h_E, W) \equiv \theta_0 = (\beta_0', \lambda_0')'$ as follows:[15]

$$\theta(\mathcal{F}_0, h_E, W) \in \underset{\tilde{\theta}}{\text{Argmin}}\ \mathbb{E}\left[\xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) h_E(z_{jt})\right]' W \mathbb{E}\left[h_E(z_{jt}) \xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta})\right].$$

If the model is well-specified ($f \in \mathcal{F}_0$) and the pseudo-true value is unique, then the pseudo-true value is the true value: $\theta_0 = \theta$. Under misspecification, θ_0 is a parameter whose value depends on $(\mathcal{F}_0, h_E, W)$. For exposition purposes, we omit this dependence in the subsequent

[15] Our definition of a pseudo-true value is closely related to the approach in White (1982) in the context of maximum likelihood. In his case, the pseudo true value minimizes the Kullback-Leibler distance between the assumed likelihood and the true likelihood, whereas in our case, the pseudo-true value minimizes a weighted sum of population moments.

analysis. Moreover, here we remain general and do not impose that W must be equal to the usual optimal weighting matrix. It is often the case in practice, that the researchers choose the identity matrix or regularize the weighting matrix.

First stage estimator $\hat{\theta}$ The first stage estimator is an empirical counterpart of the BLP pseudo-true value defined previously. The minimization is done with respect to sample analogs. Additionally, we know that there is no closed form expressions for the structural error $\xi_{jt}(f_0(.|\tilde{\lambda}), \tilde{\beta})$, and thus, we must use a feasible counterpart $\hat{\xi}_{jt}(f_0(.|\tilde{\lambda}), \tilde{\beta})$ instead.

$$\hat{\theta}(\mathcal{F}_0, h_E, \hat{W}) \equiv \hat{\theta} = \underset{\tilde{\theta}}{\text{Argmin}} \left(\sum_{j,t} \hat{\xi}_{jt}(f_0(.|\tilde{\lambda}), \tilde{\beta}) h_E(z_{jt}) \right)' \hat{W} \left(\sum_{j,t} \hat{\xi}_{jt}(f_0(.|\tilde{\lambda}), \tilde{\beta}) h_E(z_{jt}) \right). \quad (5.7)$$

The construction of the feasible structural error $\hat{\xi}_{jt}(f_0(.|\tilde{\lambda}), \tilde{\beta})$ requires the following 3 numerical approximations:

1. The econometrician does not observe a continuum of consumers as in the theoretical model but only empirical averages $\hat{s}_{jt}$ over the n_t individuals in market t.

$$\hat{s}_{jt} = \frac{1}{n_t} \sum_{i=1}^{n_t} y_{ijt}, \quad (5.8)$$

   where $y_{ijt} \in \{0; 1\}$ are i.i.d. choices over the $i = 1, \ldots, n_t$.

2. There is no closed form for $\rho_j(., x_{2t}, f_0(\cdot|\tilde{\lambda}))$, the integral has to be computed through numerical integration. A prominent example is Monte Carlo integration:

$$\hat{\rho}_j(\delta, x_{2t}, f_0(|\tilde{\lambda})) = \frac{1}{R} \sum_{r=1}^{R} \frac{\exp(\delta_j + x_{2jt} v_r)}{1 + \sum_{k=1}^{J_t} \exp(\delta_k + x'_{2kt} v_r)}, \quad (5.9)$$

   with v_r iid draws from $f_0(\cdot|\tilde{\lambda})$.

3. There is no analytical way to recover the inverse of the demand functions $\rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$. The most popular way to derive the inverse demand is by solving the following contraction:

$$C : (\cdot, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) : \delta \mapsto \delta + \log(s_t) - \log(\rho(\delta, x_{2t}, f_0(\cdot|\tilde{\lambda}))).$$

   This solution has given rise to the popular nested fixed point GMM procedure.[16]

[16] Another solution that has gained traction in the literature is the MPEC procedure (Dubé et al. (2012)) that replaces the BLP inversion at each step of the minimization by imposing equilibrium constraints on the minimization program.

In Section 5.3, we explicitly state the assumptions that allow us to neglect these approximations asymptotically.

5.2 Test procedure

Under Assumption A, and assuming $h_E(z_{jt})$ and W are such that the pseudo-true value θ_0 is unique, the following equivalence holds:

$$\begin{aligned} H_0 : f \in \mathcal{F}_0 &\iff \overline{H}_0 : (f, \beta) = (f_0(\cdot|\lambda_0), \beta_0) \\ &\iff \mathbb{E}[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)|z_{jt}] = 0 \ \ a.s.. \end{aligned}$$

The pseudo true value reduces the dimensionality of the problem by allowing us to move from a composite hypothesis $H_0 : f \in \mathcal{F}_0$ to the simple hypothesis $\overline{H}_0 : (f, \beta) = (f_0(\cdot|\lambda_0), \beta_0)$ studied previously. As we did in section 2, we propose a moment-based test of H_0.[17] Under H_0, for every set of testing instruments $h_D(z_{jt})$, the following moment conditions must hold:

$$H_0 : f \in \mathcal{F}_0 \iff \overline{H}_0 : (f, \beta) = (f_0(\cdot|\lambda_0), \beta_0) \implies \overline{H}_0' : \mathbb{E}\left[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_D(z_{jt})\right] = 0.$$

We now develop a procedure to test $\overline{H}_0'$. In comparison to the test in section 3.1, we must now account for the fact that the pseudo-true value needs to be estimated to derive the test statistic, which generates parameter uncertainty. Moreover, we propose a rigorous treatment of the numerical approximations involved in the derivation of the structural error.

Test statistic For any choice of testing instruments $h_D(z_{jt})$, our objective is to test $\overline{H}_0' : \mathbb{E}[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_D(z_{jt})] = 0$ where $\theta_0 = (\beta_0', \lambda_0')'$ is the pseudo-true value associated with

[17] Other testing approaches could have been considered. First, one could use the previous equivalence to directly test H_0 via an integrated conditional moment test. We do not follow this route for at least two reasons. First, this test will contain no information on the nature of the misspecification (it could be completely unrelated to the distribution of RC). Second, in practice the dimension of z_{jt} is often very large, which substantially reduces the power of this kind of test. Another testing approach would have entailed testing $H_0 : f \in \mathcal{F}_0$ against a larger class of densities that encompasses $\mathcal{F}_0$. For instance, if $\mathcal{F}_0$ is the family of normal distributions, encompassing families are mixtures of Gaussians with a larger number of components. We do not follow this route for two reasons. First, it is not desirable to restrict the alternative to a class of distributions that encompass the null as the econometrician does not know a priori the misspecification. Second, estimating the BLP model with a more flexible parametrization is challenging. An advantage of our test procedure is that it doesn't require estimating the model with a more flexible parametrization.

the parametric family $\mathcal{F}_0$.[18] In order to test $\overline{H}_0$, we consider the following Wald test statistic:

$$S_T(h_D, \mathcal{F}_0, \hat{\theta}) = TJ \left(\frac{1}{TJ} \sum_{j,t} \widehat{\xi}_{jt}(f_0(\cdot|\hat{\lambda}), \hat{\beta}) h_D(z_{jt}) \right)' \hat{\Sigma} \left(\frac{1}{TJ} \sum_{j,t} \widehat{\xi}_{jt}(f_0(\cdot|\hat{\lambda}), \hat{\beta}) h_D(z_{jt}) \right).$$

where $\hat{\Sigma}$ is a weighting matrix chosen by the econometrician and $\hat{\theta} = (\hat{\beta}, \hat{\lambda})$ is a consistent estimator of θ_0. The number of markets T is the dimension that we let grow to infinity to the asymptotic properties of our test. We motivate this choice in Appendix C.3. Under some regularity conditions that we make explicit in the following section, the asymptotic distribution of the test statistic under $\overline{H}_0'$ is as follows:

$$S_T(h_D, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} Z'\Sigma Z, \tag{5.10}$$

$$\text{with } \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{J} \widehat{\xi}_{jt}(f_0(\cdot|\hat{\lambda}), \hat{\beta}) h_D(z_{jt}) \xrightarrow{d} Z \sim \mathcal{N}(0, \tilde{\Omega}_0). \tag{5.11}$$

Σ is the probability limit of $\hat{\Sigma}$. We make $\tilde{\Omega}_0$ explicit in the next subsection (in particular, the derivation of $\tilde{\Omega}_0$ takes into account parameter uncertainty). Given that $\hat{\Sigma}$ is chosen by the econometrician and it is possible to derive a consistent estimator of $\tilde{\Omega}_0$, the econometrician can always simulate the asymptotic distribution of the test statistic. In some polar cases, which we present hereafter, the asymptotic distribution of our test statistic is pivotal chi-square distribution that does not require to be simulated.

Two polar cases For the sake of exposition, let us now describe two polar cases where the asymptotic distributions are pivotal chi-square distributions, which do not require to be simulated. Denote by $|\cdot|_0$ the counting norm.

1. **Sargan-Hansen J test:** If the set of estimation instruments and the set of testing instruments are the same ($h_E = h_D$), if $\hat{W}$ is the 2-step GMM optimal weighting matrix and if $\hat{\Sigma} = \hat{W}^{-1}$, then our test boils down to the usual Sargan-Hansen J test and we have under $\overline{H}_0'$:

$$S_T(h_D, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} \chi^2_{|h_E|_0 - |\theta|_0}.$$

[18] Remember that under an alternative specification, the pseudo true value also depends on the estimation instruments $h_E(z_{jt})$ and the weighting matrix.

2. **Non-redundant h_D and h_E:** if $\tilde{\Omega}_0$ has full rank and if the econometrician sets $\hat{\Sigma} = \hat{\tilde{\Omega}}_0^{-1}$, then our test statistic has the following asymptotic distribution under $\overline{H}_0'$:[19]

$$S_T(h_D, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} \chi^2_{|h_D|_0}.$$

Choice of the testing instruments As previously indicated, the power properties of our test hinge critically on the choice of testing instruments $h_D(z_{jt})$. We established that the MPI and its feasible counterparts, the interval instruments, feature attractive properties in testing $\overline{H}_0 : (f, \beta) = (f_0(\cdot|\lambda_0), \beta_0)$ against any fixed alternative. Thus, it is natural to use these instruments for the specification test above. In particular, we show that the consistency of the test with the MPI carries over to the general specification test above in Appendix B.5.

5.3 Asymptotic validity

We now study the asymptotic properties of our test when the number of markets T goes to infinity. To establish the asymptotic validity and consistency of our test, we exploit classical results on the asymptotic normality of the non-linear GMM estimator (Hansen (1982), Newey (1990)) as well on the large-T asymptotics of the BLP estimator (Freyberger (2015)). The main challenge here is to control the magnitude of the approximations that intervene in the derivation of the structural error so that they can be neglected asymptotically. Contrary to Freyberger (2015), we do not assume the convergence of any moments ex-ante and we allow for the approximation error between demand and observed market shares to be non-zero.

Assumption B

(i) $(s_t, x_t, z_t)_{t=1}^T$ *are i.i.d. across markets and are consistent with the probability model defined by equations (2.1), (2.2) and (2.3) evaluated at* (f, β);

(ii) Strong Exogeneity: $\mathbb{E}[\xi_{jt}(f, \beta)|z_{jt}] = 0$ *a.s.;*

(iii) Finite moment conditions: x_{2t} *has bounded support and* x_{1t} *has finite 4th moments.*

In B(i), we assume that the data are i.i.d. across markets, an assumption which we could relax slightly (technically, only certain moments need to be identical across markets), and that the data are generated by the BLP demand model at a given pair (f, β). In B(ii), we assume

[19] If Ω_0 is singular, one can always use directly the asymptotic distribution in 5.10 or apply the singularity-robust procedure proposed in Andrews and Guggenberger (2019).

exogeneity of our instrumental variables. Let us stress that to show the asymptotic validity of our specification test, we do not require (f, β) to be non-parametrically identified, as we just need parametric identification under H_0. In particular, we do not need all the assumptions in A. B(iii) is a necessary condition to recover the asymptotic normality of the BLP estimator. x_{1t} having finite 4th moments is standard. x_{2t} having bounded support has two purposes. First, it implies that the structural error has a finite 4th moment, Compiani (2018) makes the same assumption on price for this purpose. Second, it ensures that the mapping used in the nested fixed point algorithm is a proper smooth contraction, which allows us to prove that the NFP algorithm converges (without truncating the contraction mapping as in Berry (1994) and Berry et al. (1995)) and control for the NFP approximation bias.

Assumption C
$\mathcal{F}_0$ is such that :
(i) λ_0 belongs to the interior of Λ_0 with Λ_0 compact;
(ii) $\tilde{\lambda} \mapsto \rho(\delta, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ is well defined and continuously differentiable on Λ_0.

In C(i), we assume that, for any given DGP, the associated pseudo-true-value λ_0 associated with the family $\mathcal{F}_0$ lies in a compact space Λ_0. This condition is standard in establishing the consistency and asymptotic normality of M-estimators. Second, in C(ii), we impose that the demand function and its derivative with respect to λ should both be well defined and continuous.

Next, we impose conditions on the instruments that are used for estimation $h_E(z_{jt})$ and for testing $h_D(z_{jt})$ and on the BLP estimator itself.

Assumption D
For a given $\mathcal{F}_0$ that satisfies Assumption C and for some weighting matrix W and Σ, the following conditions must hold:
(i) Finite moments for instruments: $h_E(z_{jt})$ and $h_D(z_{jt})$ are not perfectly colinear and have finite 4th moments;
(ii) Global identification of θ_0: $\exists! \theta_0$ such that $\forall \tilde{\theta} \neq \theta_0$:

$$\mathbb{E}\left[\sum_j \xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) h_E(z_{jt})'\right] W \mathbb{E}\left[\sum_j h_E(z_{jt}) \xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta})\right] > \mathbb{E}\left[\sum_j \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_E(z_{jt})'\right] W \mathbb{E}\left[\sum_j h_E(z_{jt}) \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)\right];$$

(iii) Local identification: $\Gamma(\mathcal{F}_0, \theta_0, h_E) = \mathbb{E}\left[\sum_j h_E(z_{jt}) \frac{\partial \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)}{\partial \theta'}\right]$ *and* $\Gamma(\mathcal{F}_0, \theta_0, h_D)$ *have full column rank;*
(iv) W *and* Σ *are symmetric positive definite and* $\hat{W} \xrightarrow{P} W$, $\hat{\Sigma} \xrightarrow{P} \Sigma$*;*
(v) $\hat{\theta}$ *minimizes objective function (5.7) and satisfies the FOC of the minimization problem:*

$$\left(\sum_{j,t} \frac{\partial \hat{\xi}_{jt}(f(\cdot|\hat{\lambda}), \hat{\beta})}{\partial \theta} h_E(z_{jt})\right)' \hat{W} \left(\sum_{j,t} \hat{\xi}_{jt}(f(\cdot|\hat{\lambda}), \hat{\beta}) h_E(z_{jt})\right) = 0.$$

Assumption D restricts the class of instruments which can be used for estimation and for testing. More specifically, D(i) and D(iii) are common regularity conditions necessary to establish asymptotic results whereas D(ii) is an identification condition which ensures that the pseudo true value θ_0 is uniquely defined, which is critical to show the consistency of the BLP estimator. Finally, Assumptions D(iv) and D(v) impose regularity conditions on the weighting matrix as well as on the BLP estimator itself.

The next assumptions ensure that the numerical approximations involved in the derivation of the structural error do not interfere with the asymptotic theory.

Assumption E
(i) Let n_t *be the number of individuals in market* t, $(n_t)_{t=1}^T$ *is i.i.d. and independent from all other variables. First, it must be that* $\forall t \ \sqrt{T}\mathbb{E}(n_t^{-1/2}) \underset{T\to+\infty}{\to} 0$*. Second, observed market share* $\hat{s}_t$ *in market* t *must write:*

$$\hat{s}_{jt} = \frac{1}{n_t} \sum_{i=1}^{n_t} y_{ijt},$$

with $(y_{ijt})_{i=1}^{n_t}$ *i.i.d. draws generated by the BLP demand model at a given pair* (f, β) *conditional on* (x_t, ξ_t)*.*
(ii) Let R *be the number of simulations, then the simulated demand for product* j *writes:*

$$\hat{\rho}_{jt}(\delta, x_{2t}, f_0(\cdot|\tilde{\lambda})) = \frac{1}{R} \sum_r \frac{\exp(\delta_j + x'_{2jt} v_r)}{1 + \sum_k \exp(\delta_k + x'_{2kt} v_r)},$$

where $v_r \overset{iid}{\sim} f_0(\cdot|\tilde{\lambda})$*, and* $\frac{T}{R} \underset{T\to+\infty}{\to} 0$*.*
(iii) Let H *be the stopping time for the contraction (which depends on* T*) and* ϵ *the fixed Lipschitz constant of the contraction mapping used to invert the demand function, then it must be that* $\sqrt{T}\epsilon^H \underset{T\to+\infty}{\to} 0$*.*

A sufficient condition for E(i) to hold is that the minimum number of individuals observed in any market is of higher order than the total number of markets. This condition can be checked in practice.[20] Assumptions E(ii) and E(iii) can also be checked in practice and are more manageable because R and H are chosen by the researcher and can always be increased so that these assumptions hold.

Given our assumptions, we derive the asymptotic distribution of our test statistic under the null, and show that the test is consistent.

Theorem 5.1 *Let $\hat{\theta} = \hat{\theta}(\mathcal{F}_0, \hat{W}, h_E)$ be the BLP estimator associated with distributional assumption $\mathcal{F}_0$, weighting matrix $\hat{W}$, estimating instruments h_E. Under assumptions B-E,*

- *Under $\bar{H}_0'$:* $\mathbb{E}\left[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_D(z_{jt})\right] = 0$,

$$S_T(h_D, \mathcal{F}_0, \hat{\theta}) \underset{T\to+\infty}{\overset{d}{\to}} Z'\Sigma Z, \qquad Z \sim \mathcal{N}(0, \tilde{\Omega}_0),$$

$$\text{where} \quad \tilde{\Omega}_0 = \begin{pmatrix} I_{|h_D|_0} & G \end{pmatrix} \begin{pmatrix} \Omega(\mathcal{F}_0, h_D) & \Omega(\mathcal{F}_0, h_D, h_E) \\ \Omega(\mathcal{F}_0, h_D, h_E)' & \Omega(\mathcal{F}_0, h_E) \end{pmatrix} \begin{pmatrix} I_{|h_D|_0} \\ G' \end{pmatrix},$$

$$\Omega(\mathcal{F}_0, h_D, h_E) = cov\left(\sum_j \xi_{jt}(f(.|\lambda_0), \beta_0) h_D(z_{jt}), \sum_j \xi_{jt}(f(.|\lambda_0), \beta_0) h_E(z_{jt}) \right),$$

$$G = -\Gamma(\mathcal{F}_0, \theta_0, h_D) \left[\Gamma(\mathcal{F}_0, \theta_0, h_E)' W \Gamma(\mathcal{F}_0, \theta_0, h_E)\right]^{-1} \Gamma(\mathcal{F}_0, \theta_0, h_E)' W.$$

- *Under H_a' :* $\mathbb{E}\left[h_D(z_{jt}) \xi_{jt}(f_0(.|\lambda_0), \beta_0)\right] \neq 0$,

$$\forall q \in \mathbb{R}^+, \ \mathbb{P}(S_T(h_D, \mathcal{F}_0, \hat{\theta}) > q) \underset{T\to+\infty}{\to} 1.$$

The proof of Theorem 5.1 is in Appendix B.4 and comprises three main steps. First, we show that under the assumptions in E, the numerical approximation becomes asymptotically negligible. Second, we show the consistency and asymptotic normality of the BLP estimator. Finally, we derive the asymptotic distribution of the test statistic, taking into account parameter uncertainty (θ_0 is estimated and not observed). The apparent complexity of the asymptotic variance-covariance matrix Ω_0 is a direct consequence of parameter uncertainty.

[20] Note that by making stronger assumptions on the higher moments and the support of the observed characteristics, it is possible to find milder conditions on the number of individuals relative to the number of markets.

6 Monte Carlo experiments

In this section, we conduct three distinct sets of Monte Carlo experiments. First, we implement a simple simulation exercise to assess the effects of incorrectly specifying the distribution of random coefficients on quantities of interest such as price elasticities or cross-price elasticities, which are known to play a key role in shaping the counterfactuals. In a second set of Monte Carlo experiments, we study the finite sample performances of the specification test developed in section 5 with different sets of testing instruments. We first examine the size of our test in finite sample. Then, we investigate the power properties of our test under alternative specifications (with alternatives including Gaussian mixtures, gamma distributions and local alternatives). We show that our test with the interval instruments significantly outperforms the traditional J-test with the usual instruments. Finally, in the last Monte Carlo exercise, we study the performance of the interval instruments to estimate the parameters of the model by means of comparison with the commonly used instruments in the literature.

6.1 Simulation design

For the sake of exposition, we will keep the same simulation design for all the simulation experiments. The simulation design closely follows the simulation design used in Dubé et al. (2012), Reynaert and Verboven (2014). The market includes $J = 12$ products, which are characterized by 3 exogenous product attributes x_a, x_b and x_c that follow a joint normal distribution. The price p is endogenous and depends on the observed and unobserved characteristics and on some cost shifters c_1 and c_2. Consumer heterogeneity is present only in x_c, and the random coefficient v_i associated with x_c follows various distributions depending on the simulation exercise. The sample size T varies between 50, 100 and 200 markets. We can summarize the DGP as follows:

$$u_{ijt} = 2 + x_{ajt} + 1.5x_{bjt} - 2p_{jt} + x_{cjt}v_i + \xi_{jt} + \varepsilon_{ijt} \qquad \xi_{jt} \sim \mathcal{N}(0,1),\ \varepsilon_{ijt} \sim EV1,$$

$$\text{and} \begin{bmatrix} x_{a,j} \\ x_{b,j} \\ x_{c,j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 & 0.3 \\ -0.8 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} \right),$$

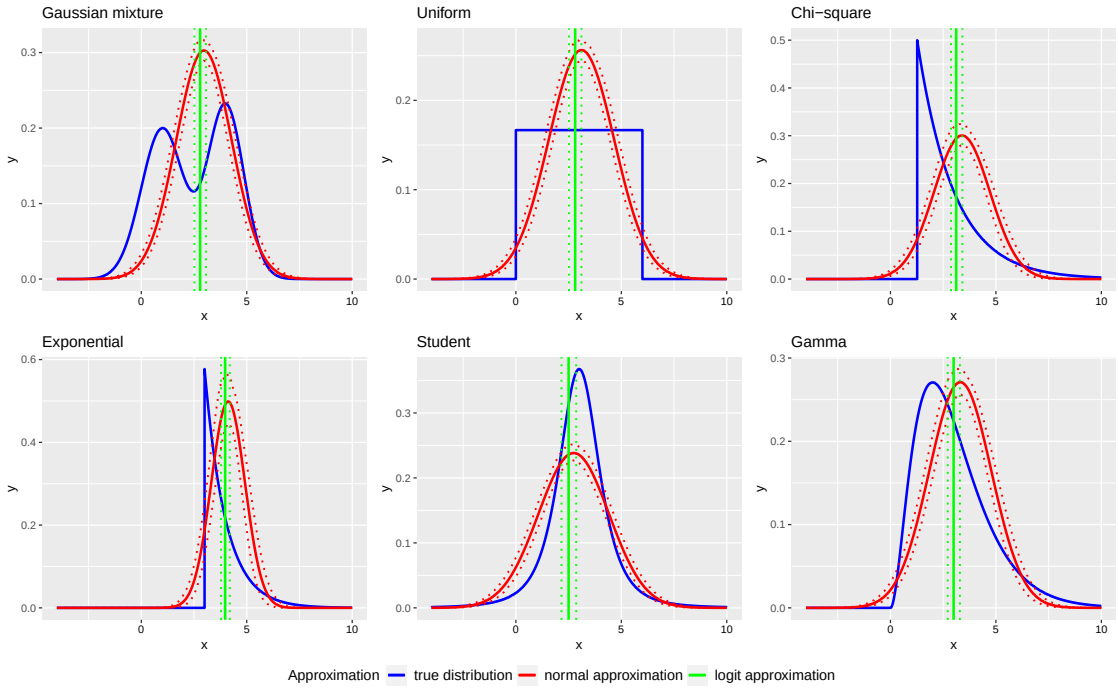
$$p_{jt} = 1 + \xi_{jt} + u_{jt} + \sum_{k=a}^{c} x_{kjt} + c_{1jt} + c_{2jt} \quad \text{with } u_{j,t} \sim U[-4,-2],\ c_{1jt} \sim U[2,4] \text{ and } c_{2jt} \sim U[3,5].$$

Market shares are generated by integrating over $20,000$ consumers. This allows us to essentially remove the approximation error between the observed and theoretical market shares.

6.2 Counterfactuals under an alternative distribution

We now present a simple exercise to illustrate how the misspecification of random coefficients can affect the estimation of quantities of interest such as price elasticities and cross-price elasticities. To do so, we simulate data using the simulation design introduced above and we take various distributions for the random coefficient v_i (respectively: Gaussian mixture, Uniform, Chi-square, Exponential, Student, Gamma). We ensure that all the distributions have the same mean and variance (3 and 3, respectively). For each distribution, we simulate $T = 100$ markets of data and we estimate the model either assuming no heterogeneity (simple logit) or assuming that v_i is normally distributed. We replicate the same exercise 500 times for each distribution. This allows us to recover the mean estimate for the parameters as well as to construct 95% "confidence bins" (by trimming the observations below the 2.5% quantile and above the 97.5% quantile). We plot the true densities and their estimated counterparts under the normal and logit assumptions in Figure 1. We observe that the estimated logit parameters and the estimated means of the normal distributions always coincide and are close to 3 for all the distributions. However, there is some variation between the different specifications. For instance, the estimated means are larger with the exponential distribution. The estimated variances also vary from one specification to the other. The estimated variances for the exponential distribution are smaller, while they are larger for the student distribution.

Figure 1: True densities and estimated densities under normal and logit specifications



In a second stage, we simulate $N = 5,000$ draws from the true distributions as well as from the estimated logit and normal approximations to compute the demand, the price-elasticity and the cross-price elasticity for the product j^* with the highest value for x_c.[21] The cross-price elasticity is arbitrarily taken for product $j = 1$ with respect to p_{j^*}. We derive the quantities of interest for 100 equally spaced values of p_{j^*} ranging in $]0, 10[$. We plot the elasticities in Figure 2 and cross-price elasticities in Figure 3 generated by the true distribution as well as those generated by the logit and normal approximations, respectively. We proceed similarly with the demand functions. We see in Figure 9 in Appendix).

One can observe that, as expected, the logit specification poorly replicates the substitution patterns. In particular, it consistently overstates the magnitude of the elasticities and cross-elasticities with respect to the true ones. The absence of consumer heterogeneity on characteristic c implies that consumers can "renounce' more easily to product j^* when its price increases. By introducing some heterogeneity, the normal approximation somewhat attenuates this issue. However, significant discrepancies in the shape of elasticities and cross-price elasticities remain. As most counterfactual analyzes rely on the substitution patterns

[21] The expressions for both price-elasticities and the cross-price elasticities are in Appendix D.1.

generated by the model, these differences will inevitably create significant biases.

Figure 2: Price elasticities

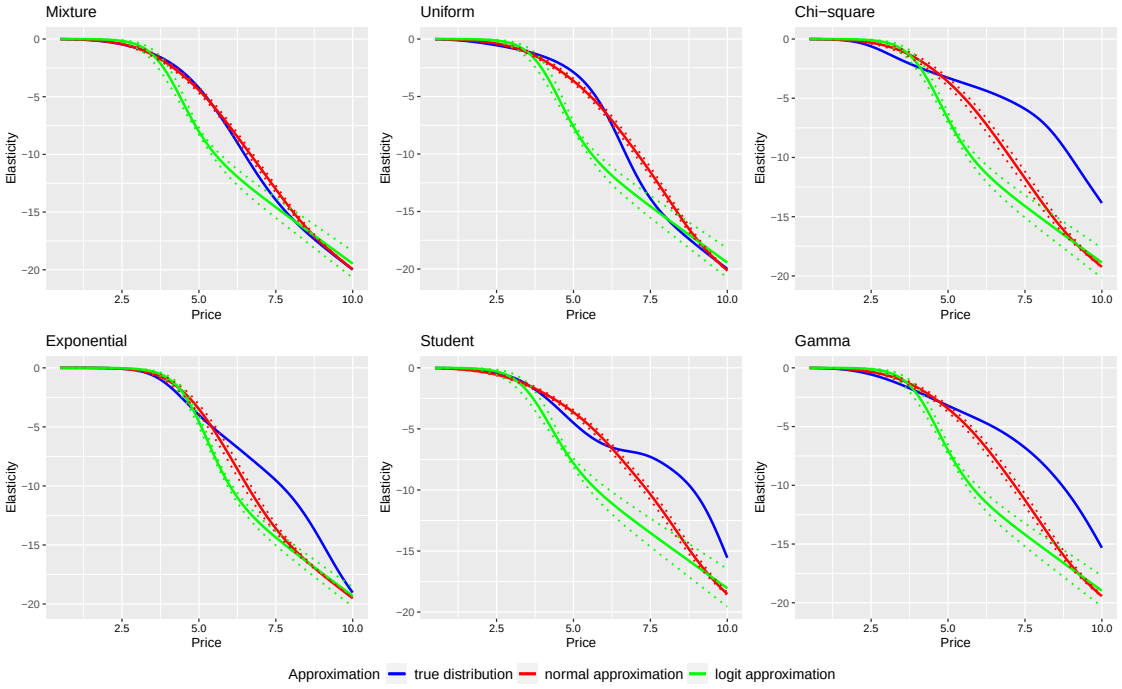
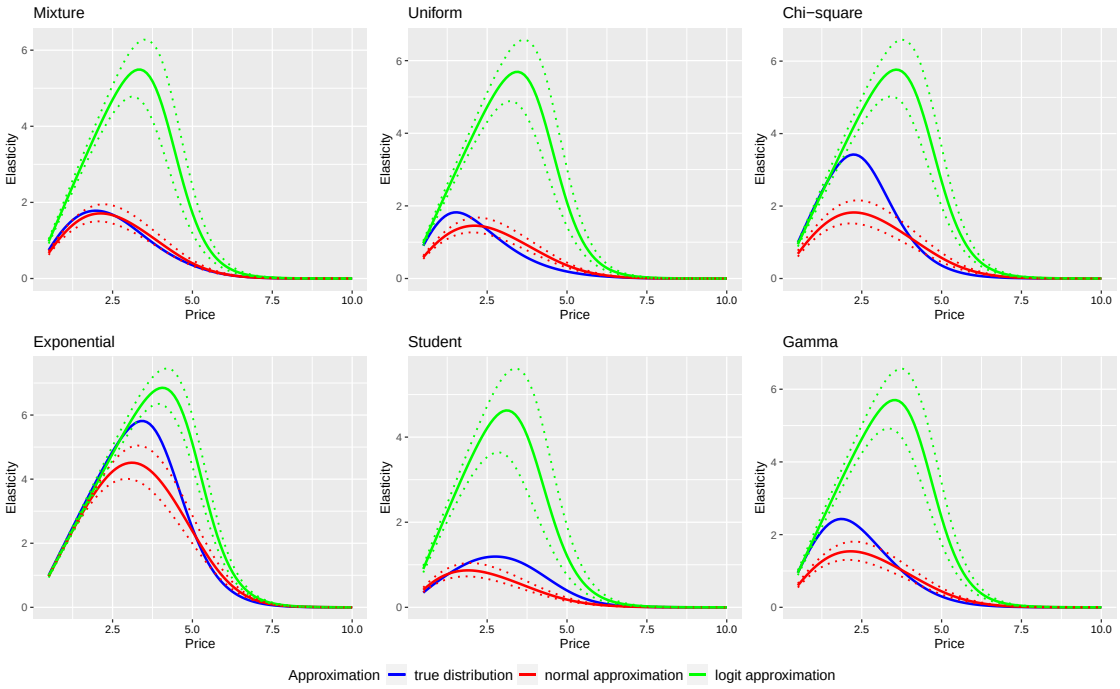


Figure 3: Cross-price elasticities



6.3 Finite sample performance of the specification test

We now study the empirical size and power of our test under different sample sizes and for different sets of testing and estimating instruments. Once again, the data are generated according to the simulation design exhibited previously for various distributions of v_i. The assumption made throughout the simulations is $H_0 : f \in \mathcal{F}_0$, where $\mathcal{F}_0$ is the family of normal distributions. In other words, we always assume that the random coefficient is normally distributed and we test this hypothesis. We set the nominal size to 5%. We study the finite sample performances of the specification test that we presented in section 5 using different sets of estimation and testing instruments. For estimation, we take the instruments commonly adopted by practitioners: the differentiation instruments of Gandhi and Houde (2019) and the "optimal" instruments of Reynaert and Verboven (2014). Both of these sets are approximations of the classical optimal instruments. Second, we compare the performance of the test when performing the standard Sargan-Hansen J test (i.e. when we use the same instruments for testing and estimation) and when we use the global and local approximations of the MPI that we constructed in sections 4.2 and 4.1. We denote the latter tests as I Local and I Global respectively. The BLP estimator is computed following the NFP GMM procedure described

in section 5.1. For the optimization, only an analytic Jacobian is provided. We ensure that the number of tested restrictions is of the same magnitude across the different sets of instruments. More details on the exact sets of instruments and on the estimation procedure for this specific set of simulations are given in Appendix D.2.

6.3.1 Empirical size

The size is the probability of rejecting the null hypothesis when the null is true, so we compute the empirical size by counting and averaging the number of times we reject the null for nominal size 5% over the 1,000 simulations when the random coefficient v_i is normally distributed. Below in Table 1, we report the empirical sizes of the test with the different sets of instruments described above for the different sample sizes $T \in \{50, 100, 200\}$ and for different distributions of the RC such that $v_i \sim f \in \mathcal{F}_0$.

Table 1: Empirical size for nominal size 5% (1000 replications)

Number of markets	T=50						T=100						T=200					
Estimation instruments	"Differentiation"			"Optimal'			"Differentiation"			"Optimal'			"Differentiation"			"Optimal'		
Test type	J	I Global	I local	J	I Global	I local	J	I Global	I local	J	I Global	I local	J	I Global	I local	J	I Global	I local
$v_i \sim \mathcal{N}(-1, 0.5^2)$	0.294	0.083	0.091	0.145	0.078	0.063	0.138	0.078	0.058	0.094	0.084	0.047	0.08	0.052	0.053	0.064	0.05	0.04
$v_i \sim \mathcal{N}(0, 0.75^2)$	0.293	0.084	0.085	0.148	0.081	0.071	0.137	0.061	0.06	0.1	0.059	0.05	0.074	0.053	0.045	0.062	0.048	0.036
$v_i \sim \mathcal{N}(1, 1^2)$	0.287	0.084	0.083	0.142	0.084	0.073	0.142	0.055	0.054	0.098	0.053	0.047	0.079	0.042	0.03	0.058	0.035	0.025
$v_i \sim \mathcal{N}(2, 2^2)$	0.288	0.087	0.077	0.145	0.071	0.072	0.138	0.069	0.051	0.099	0.053	0.056	0.077	0.044	0.041	0.069	0.037	0.044
$v_i \sim \mathcal{N}(3, 3^2)$	0.287	0.089	0.071	0.137	0.075	0.066	0.145	0.074	0.06	0.098	0.06	0.061	0.076	0.044	0.037	0.061	0.046	0.046

We observe that with a moderate sample size ($T = 50, J = 12$), all the tests are over-sized. This is within expectations and due to the approximations inherent to the estimation of the BLP models as described in section 5 and the relatively large number of instruments used for estimation and testing purposes.[22] However, we notice that the Sargan-Hansen J tests are much more over-sized than the tests with the interval instruments: the rejection rate is above 25% for the Sargan-Hansen J test with differentiation instruments vs around 8% for the I test. Increasing the sample size improves the tests' empirical levels and shifts them towards the nominal level, which is a good indication of the validity of our test. Even with

[22]The number of over-identifying restrictions lies between 6 and 8. The Sargan-Hansen J tests are known to suffer from size distortions as the number of instruments increases.

a relatively large number of markets ($T = 200$), the Sargan-Hansen J tests remain slightly oversized (rejection rate is still slightly above 5%). On the other hand, for the test with interval instruments, the empirical size appears to match the nominal level for all but two configurations, where it seems to be slightly undersized.

6.3.2 Empirical power

Power is the probability of rejecting the null hypothesis under an alternative. We compute the empirical power by counting and averaging the number of times we reject the null for the test of nominal size 5% over the 1000 simulations when the distribution of random coefficients is misspecified. The simulation setup remains the same as previously with the only modification being that the true distribution of v_i is now either a mixture of normals or a Gamma. We report the power against the different alternatives in the subsequent tables. The main takeaway from our results is that the test with the interval instruments as testing instruments (I global and I local) largely outperforms the traditional Sargan-Hansen J-test against all the alternative distributions considered in our simulations.

Power against Gaussian mixture alternatives We simulate data with the random coefficients distributed according to the Gaussian mixtures described below. We plot the true distributions in Figure 4. We report the results in Table 2. We observe that the test with the interval instruments has great power against all the mixtures tested. The rejection rates go to 1 very quickly in comparison to the Sargan-Hansen J tests.

$$
\begin{aligned}
&v = Dv_1 + (1-D)v_2, \quad \mathbb{P}(D=1) = p, \quad \mathbb{P}(D=0) = 1-p, \\
&v_1 \sim \mathcal{N}\left(-\sqrt{\frac{3p}{1-p}} + 2, 1\right) \quad v_2 \sim \mathcal{N}\left(\sqrt{\frac{3(1-p)}{p}} + 2, 1\right),
\end{aligned}
$$

with $p \in \{0.1; 0.2; 0.3; 0.4; 0.5\}$.

Figure 4: Densities, Gaussian mixture alternatives

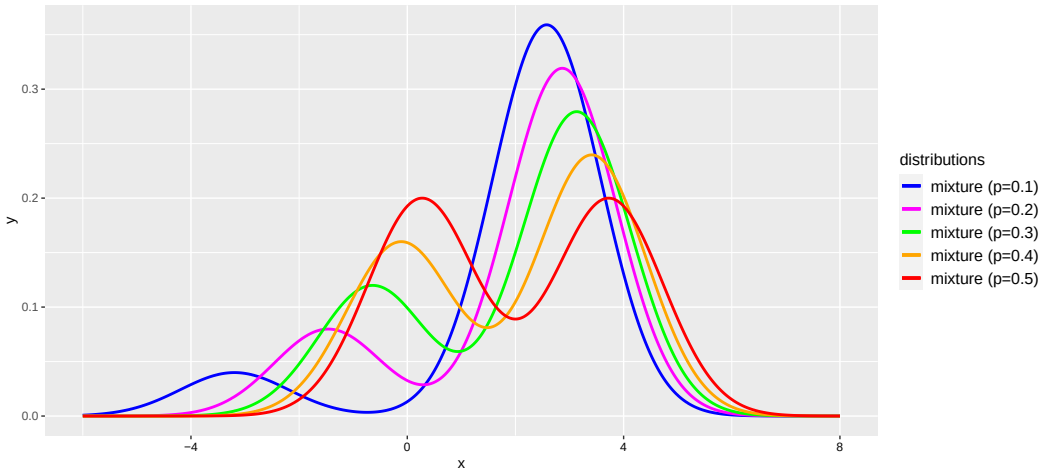


Table 2: Empirical power, Gaussian mixture alternatives (1000 replications)

Number of markets	T=50						T=100						T=200					
Estimation instruments	"Differentiation"			"Optimal'			"Differentiation"			"Optimal'			"Differentiation"			"Optimal'		
Test type	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local
Mixture 1	0.533	0.991	0.987	0.719	0.989	0.989	0.604	1	1	0.967	1	1	0.829	1	1	1	1	1
Mixture 2	0.626	0.996	0.998	0.613	0.997	0.998	0.723	1	1	0.905	1	1	0.933	1	1	1	1	1
Mixture 3	0.629	0.992	0.995	0.43	0.996	0.997	0.741	1	1	0.7	1	1	0.941	1	1	0.977	1	1
Mixture 4	0.601	0.983	0.982	0.275	0.981	0.981	0.713	1	0.999	0.368	1	1	0.921	1	1	0.672	1	1
Mixture 5	0.56	0.907	0.904	0.157	0.9	0.906	0.635	0.993	0.995	0.124	0.995	0.996	0.855	1	1	0.146	1	1

Power against Gamma alternatives We simulate data with the random coefficients distributed according to the Gamma distribution described below. We plot the true distributions in Figure 5. We report the results in table 3. We observe that the test with interval instruments has great power against all the Gamma distributions tested except for the first one, which we can see on the plot has a distribution that is relatively close to a normal distribution. Even for the first Gamma distribution, it still outperforms the traditional sets of instruments. For all the other Gamma distributions, the rejection rates go to 1 very quickly in comparison to the Sargan-Hansen J-tests. This confirms the superiority of the interval instruments in detecting misspecification in the distribution of random coefficients. In Appendix D.2, we also study

the power properties of our test against local alternatives.

$$v \sim \Gamma(2,k) \quad with\ k \in \{0.25, 0.5, 0.75, 1, 1.5\}$$

Figure 5: Densities, Gamma alternatives

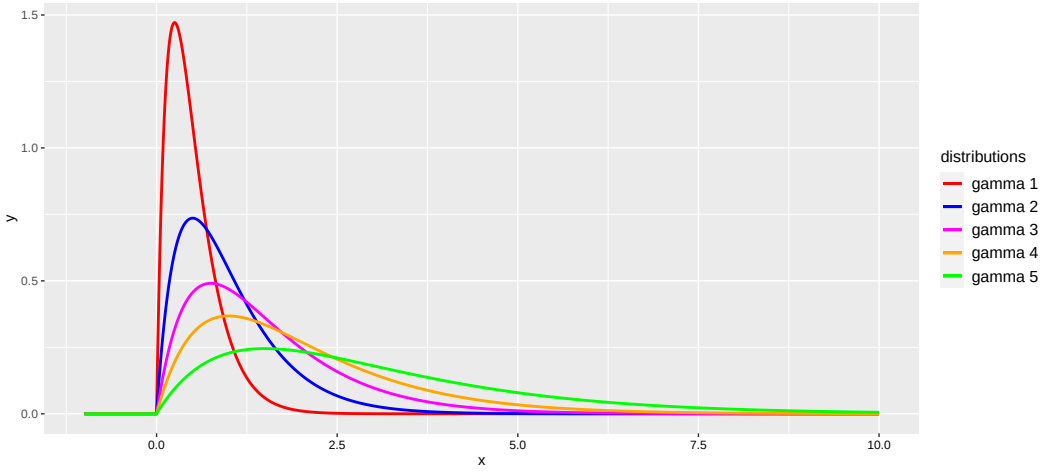


Table 3: Empirical power, Gamma alternatives (1000 replications)

Number of markets	T=50						T=100						T=200					
Estimation instruments	"Differentiation"			"Optimal'			"Differentiation"			"Optimal'			"Differentiation"			"Optimal'		
Test type	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local	J	I Global	I Local
Gamma 1	0.293	0.106	0.093	0.142	0.082	0.074	0.154	0.083	0.073	0.094	0.092	0.08	0.118	0.155	0.139	0.066	0.156	0.138
Gamma 2	0.516	0.747	0.752	0.14	0.781	0.77	0.562	0.983	0.978	0.095	0.982	0.98	0.492	1	1	0.08	1	1
Gamma 3	0.607	0.96	0.962	0.157	0.963	0.969	0.693	0.998	1	0.156	1	1	0.922	1	1	0.161	1	1
Gamma 4	0.622	0.97	0.99	0.207	0.962	0.995	0.748	0.999	1	0.263	1	1	0.933	1	1	0.412	1	1
Gamma 5	0.687	0.991	0.999	0.371	0.988	0.999	0.812	1	1	0.585	1	1	0.976	1	1	0.865	1	1

6.4 Finite sample performance of interval instruments for estimation

In our last simulation exercise, we evaluate the performance of our interval instruments in estimating the parameters associated with the RC when the distribution of random coefficients is flexibly parametrized. To do so, we simulate data with a distribution of random coefficients

following a mixture of Gaussians and we estimate the parameters of this mixture. We provide a method to estimate the parameters when the distribution of the RC is a mixture in section C.6 of the Appendix. In particular, we provide a new parametrization of the model, which yields substantial practical gains and may be of interest to researchers independent of the rest of the paper. The simulation design remains the same as previously. We assume that the random coefficient v_i is distributed according to the following mixture: $v_i \sim D_i \, \mathcal{N}(-2, 0.5) + (1 - D_i) \, \mathcal{N}(4, 0.5)$ with $\mathbb{P}(D_i = 1) = 0.25$. Thus, there are 5 parameters associated with the distribution of RC: the means and variances of each component of the mixture and the mixing probability. Our objective is to compare the performance of the global and local interval instruments with the instruments commonly used by practitioners: the differentiation instruments from Gandhi and Houde (2019) and the "optimal instruments" from Reynaert and Verboven (2014). In Table 4, we report the empirical biases and the square root of the MSE for the estimators of the non-linear parameters for each set of instruments and for the different sample sizes. In Appendix D.3, we report the same information for the linear parameters (see Tables 14, 15, and 17) as well as the distribution of the empirical distribution of the non-linear estimates. Table 4 allows us to directly compare the performances of the three sets of instruments in estimating the non-linear parameters. We first observe that for all the sets of instruments, the empirical biases and $\sqrt{MSE}$ of the estimators decrease when the sample size increases, which is reassuring. Furthermore, it appears clearly that the differentiation instruments perform worse than the "optimal instruments" and the interval instruments. The empirical $\sqrt{MSE}$ of the estimators with the differentiation instruments is up to 12 times larger than with the interval instruments and up to 6 times larger than with the "optimal instruments". We reach the same conclusions when we study empirical biases. The interval instruments appear to perform better than the "optimal instruments" even if the difference is less significant than with the differentiation instruments. For the sake of conciseness, we do not report the results obtained with a mixture of 3 components but the observations we make with two components are even more exacerbated. In Appendix D.3, as a means of comparison, we perform the same exercise when the distribution of random coefficients is a simple Gaussian and here, we do not observe any significant differences between the different sets of instruments, which confirms that the interval instruments make a difference when the distribution of RCs is flexible.

Table 4: Estimation non-linear parameters of the mixture (1000 replications)

	Instruments	Differentiation					"Optimal"					Interval Global					Interval Local				
	Parameter	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L
Sample size	true	-2	0.5	4	0.5	0.25	-2	0.5	4	0.5	0.25	-2	0.5	4	0.5	0.25	-2	0.5	4	0.5	0.25
T=50, J=12	bias	0.214	0.184	-0.022	-0.045	0.027	0.076	0.059	0.026	-0.111	0.01	0.017	0	-0.045	0.004	0.005	-0.006	-0.005	-0.039	-0.001	0.003
	$\sqrt{MSE}$	0.633	0.734	0.281	0.35	0.075	0.361	0.483	0.212	0.281	0.036	0.277	0.391	0.227	0.259	0.024	0.251	0.34	0.214	0.244	0.019
T=50, J=20	bias	0.189	0.347	0.022	-0.081	0.025	0.074	0.11	0.028	-0.089	0.01	0.013	0.042	-0.018	-0.003	0.004	0.019	0.033	-0.023	0.01	0.003
	$\sqrt{MSE}$	0.566	0.887	0.184	0.291	0.059	0.328	0.563	0.163	0.228	0.033	0.248	0.415	0.166	0.22	0.021	0.228	0.38	0.15	0.184	0.018
T=100, J=12	bias	0.233	0.226	0.02	-0.066	0.027	0.054	0.037	0.019	-0.066	0.007	0.004	-0.012	-0.027	0.005	0.002	0	0	-0.028	0.007	0.001
	$\sqrt{MSE}$	0.592	0.703	0.256	0.305	0.072	0.279	0.4	0.154	0.211	0.028	0.167	0.282	0.157	0.201	0.013	0.127	0.225	0.143	0.164	0.005
T=100, J=20	bias	0.198	0.423	0.047	-0.101	0.025	0.074	0.107	0.033	-0.074	0.01	-0.009	-0.005	-0.008	-0.009	0.001	-0.003	0.004	-0.01	0.004	0.001
	$\sqrt{MSE}$	0.552	0.89	0.164	0.27	0.055	0.311	0.52	0.129	0.194	0.034	0.115	0.264	0.115	0.169	0.005	0.104	0.226	0.103	0.125	0.004
T=200, J=12	bias	0.184	0.167	0.011	-0.049	0.019	0.026	0.011	0.021	-0.061	0.004	-0.006	-0.027	-0.015	-0.001	0.001	0.002	-0.007	-0.016	0.006	0.001
	$\sqrt{MSE}$	0.466	0.601	0.176	0.262	0.053	0.184	0.313	0.113	0.172	0.018	0.088	0.219	0.108	0.164	0.003	0.091	0.174	0.099	0.123	0.003

7 Empirical application

The objective of the empirical exercise is twofold. First, we want to verify how well our instruments perform at estimating a flexible distribution of RCs using a real data set. Second, we want to study how the shape of the distribution of RCs can modify key counterfactual quantities such as the price elasticities or the pass-through, and check whether the results we obtain are consistent with the findings in Miravete et al. (2022). To do so, we estimate demand for cars using data on new car registrations in Germany from 2012 to 2018.[23] There are many reasons to focus on the car market. First, cars are highly differentiated products, which makes the BLP framework particularly adapted to this market. As a result, the BLP demand model has been widely applied to study the car industry (e.g., Berry et al. (1995), Grigolon et al. (2018), Petrin (2002)) and one can easily compare our results with previous results obtained in the literature under different specifications. Second, there are many policy-relevant questions related to this market. In particular, the role of road transport in air pollution is significant and many countries have implemented tax policies to reduce the CO2 emissions generated by car transportation.[24] An important strand of the literature has investigated the perfor-

[23]The dataset was kindly provided to us by Kevin Remmy https://kevinremmy.com/research/.

[24]In 2017, road transport was responsible of approximately 19% of total greenhouse has emissions in EU-28 Retrieved from https://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-12 on Octo-

mance of these different taxation schemes (Alberini and Horvath (2021), Allcott and Wozny (2014), D'Haultfœuille, Givord, and Boutin (2014), Durrmeyer (2022), Durrmeyer and Samano (2018), Gillingham and Houde (2021), Grigolon et al. (2018), Huse and Koptyug (2022), Kunert (2018)). Other policy-relevant questions include the impact of import tariffs (Miravete et al. (2018)) and the determinants of market power (Berry et al. (1995), Grieco, Murry, and Yurukoglu (2022)). To answer these questions, the researcher must often estimate the demand for cars. The credibility of the implied analysis depends critically on how well the model can reproduce the underlying substitution patterns and the shape of the demand curve. To this end, it is essential to have a demand system that is sufficiently flexible, and particularly so with respect to the random coefficient on price. In this section, we use our instruments to estimate a Gaussian mixture as the random coefficient associated with price. Moreover, we use our test to assess how moving from the usual Gaussian RC to the Gaussian mixture decreases the degree of misspecification. Finally, we compare the counterfactual quantities under a Gaussian mixture and the traditional specifications (Gaussian RC and logit). In line with the findings in Miravete et al. (2022), our results indicate that the Gaussian mixture yields higher pass-through rates and curvatures.

7.1 Data

The data set includes state-level new car registrations, publicly available by the German Federal Motor Transport Authority (KBA) from 2012 to 2018. This gives us 112 markets defined by state-year pairs. Data on car characteristics and price are scraped from General German Automobile Club and include horsepower, engine type, size, weight, fuel cost, CO2 emission, number of doors, segment, and body type. The data set is at a granular level where every car is uniquely identified by its manufacturer and its type key code (HSN/TSN) that is defined according to the characteristics of the car. Following the literature, we aggregate products with the same brand, model, engine type, and body combination (e.g. BMW-1 Series-Diesel-Hatchback).[25] Likewise, we follow the literature and define the market size as the number of households in the market. To construct market shares, we simply divide new car registrations of a given product by the market size. The data set is complemented by information on demographics such as the number of households or the average income per household at the

ber 21, 2022.

[25] In aggregating the products from the HSN/TSN level, we use the characteristics of the car with the highest sales.

state-year level and yearly average gas price data from ADAC.[26]

Summary statistics Shares of products sold by engine type are presented in Table 5. We focus our analysis on combustion engine vehicles as in our sample period electric-vehicle cars constitute a small market share (always less than 5% of the sold vehicles) and can be seen as a distinct market. Between diesel and gasoline cars, we observe that the market share for diesel decreases over time, starting from 2016. The timing is in line with the emissions scandal, known as the Dieselgate, which started in September 2015.

Table 5: Shares (%) of new registrations by engine type

	Year						
Fuel Type	2012	2013	2014	2015	2016	2017	2018
Diesel	46.8	46.1	46.3	46.4	43.9	36.2	30.0
Gasoline	52.6	52.9	52.6	52.3	54.4	60.8	66.5
Battery EV	0.1	0.2	0.3	0.4	0.3	0.7	1.1
Hybrid EV	0.5	0.8	0.7	0.6	1	1.4	1.6
Plug-in hybrid EV	0	0	0.1	0.3	0.4	0.9	0.9

Table 6 provides sales-weighted averages for prices and observed characteristics. We observe that the difference in fuel consumption and resulting fuel costs steadily ranks diesel above gasoline. However, the average price of diesel cars sold is higher than gasoline cars. This implies a potential trade-off in terms of the costs of car ownership at the time of purchase. With a fixed mileage in mind, a consumer with high sensitivity to fuel costs might be willing to pay a higher price for a more fuel-efficient car. We also observe that the horsepower and the size of the newly registered cars increase over time.

[26]State level income https://ec.europa.eu/eurostat/web/products-datasets/-/nama_10r_2hhinc

Table 6: Summary Statistics (Sales weighted)

	Year						
	2012	2013	2014	2015	2016	2017	2018
Diesel							
Price/income	0.74	0.72	0.73	0.72	0.71	0.69	0.68
Size (m2)	8.31	8.31	8.32	8.36	8.42	8.48	8.53
Horsepower (kW/100)	1.09	1.07	1.11	1.11	1.14	1.16	1.21
Fuel cost (euros/100km)	7.90	7.18	6.63	5.53	4.94	5.25	5.83
Fuel cons. (Lt./100km)	5.19	4.98	4.89	4.73	4.61	4.61	4.71
CO2 emission (g/km)	136.19	130.50	127.69	123.58	120.42	120.49	123.27
Nb. of products/market	133	138	146	150	151	149	143
Gasoline							
Price/income	0.46	0.46	0.46	0.46	0.46	0.45	0.43
Size (m2)	7.23	7.27	7.28	7.30	7.36	7.46	7.53
Horsepower (kW/100)	0.79	0.78	0.80	0.82	0.85	0.88	0.91
Fuel cost (euros/100km)	9.48	8.61	8.11	7.27	6.69	7.06	7.40
Fuel cons. (Lt./100km)	5.76	5.47	5.40	5.31	5.25	5.34	5.38
CO2 emission (g/km)	135.80	128.18	125.27	122.89	121.22	122.86	123.26
Nb. of products/market	157	171	179	185	186	193	188

Note: Provided statistics are sales weighted averages across products. Total number of markets (State*Year) is 112 .

Inter-market variation Our dataset contains both geographical variation and time variation, as we observe the sales in every state in Germany over the period 2012-2018. States in Germany differ significantly in terms of income per inhabitant, population density and average distance driven.[27] It is fundamental to take this inter-market variation into account in our empirical specification for two reasons. First, our model postulates that consumers' preferences are the same across markets. However, we observe that the market shares vary from one state

[27]For the population density 2019 (inh/km^2): 69 (Mecklenburg-Vorpommern) to 4118 (Berlin) (from Federal Statistical Office of Germany (Destatis)), GDP per capita 2019: 28.9k (Mecklenburg-Vorpommern) to 67k (Hamburg) (retrieved from https://www.ceicdata.com/en/germany/esa-2010-gdp-per-capita-by-region/gdp-per-capita-bremen on 05 November 2022). For average driving distance in 2019: 13079 km (Mecklenburg-Vorpommern) to 9531 (Berlin) retrieved from https://de.statista.com/statistik/daten/studie/644381/umfrage/fahrleistung-privater-pkw-in-deutschland-nach-bundesland/ on 19 September 2022.

to the other even if the choice set remains the same. This feature of the data can only be explained if we let the preferences vary from one market to the other. Second, in section 2.3, we saw that there needs to be sufficient variation in the product characteristics across markets to identify the distribution of RCs. By interacting product characteristics with state demographics, we achieve both objectives: we shift the preferences to a more common representation and we create variation in the product characteristics across markets. To choose which interaction terms to include in the utility function, we first create market specific sales-weighted characteristics for the following variables: price, fuel cost, size, horsepower, height, gasoline dummy, and foreign dummy (equal to one if the manufacturer of the car is not German). Then, we regress these quantities on the demographics of interest: average income, population density, and a time trend. Last, we select the interaction terms that explain the best the variation in sales-weighted characteristics (namely, the variables with a p-value below $1e^{-10}$). The results of these regressions are presented in Table 7. They suggest that income shifts positively the preferences for price, size, and horsepower (i.e. higher income is associated with larger cars, and higher horsepower). In contrast, income shifts negatively the preferences for foreign status, height, and gasoline status.[28] Although weaker, a similar pattern is observed for the effect of population density on car characteristics.

Table 7: Linear regressions of sales-weighted car characteristics on demographic characteristics

	Income(/1000)	Population density (/100)	Time trend
Price($\times$1000)	0.138**	0.069*	0.286*
	(0.013)	(0.011)	(0.059)
Fuel cost (euros/100km)	-0.0069	-0.0036	0.3587**
	(0.0063)	(0.0056)	(0.0296)
Size(m^2)	0.0058**	0.0018*	0.0176*
	(0.00079)	(0.00070)	(0.00371)
Horsepower (KW/100)	0.0028**	0.0012*	0.0129**
	(0.00028)	(0.00025)	(0.00132)
Foreign	-0.0050**	-0.0014*	0.0295**
	(0.00052)	(0.00046)	(0.00246)
Height(m)	-0.00051**	-0.00043**	0.00181*
	(0.000061)	(0.000054)	(0.000286)
Gasoline	-0.0067**	-0.0024*	0.0131*
	(0.00059)	(0.00053)	(0.00280)

Note: * p-value lower than 0.01, ** p-value lower than $1e^{-10}$.

[28] In the main analysis, we use price/income to capture the income effect.

Instruments for the endogeneity of price To instrument for price, we use a combination of variables on the intensity of competition and cost shifters. To measure the intensity of competition, we consider the number of competing products of the same class and engine type in a given market, and the number of competing products of the same engine type in a given market. As for cost shifters, we use three complementary datasets: the mean hourly labor cost, the price of steel (interacted with the weight of the car), and exchange rates between Germany and the country of assembly.

1. Labor cost: we use the mean nominal hourly labor cost per employee in the manufacturing sector of the country of assembly of the models. The data on labor costs come from International Labor Organization Statistics (ILOSTAT).[29]

2. Price of steel: we collect the price of steel futures in January of each year.

3. Exchange rates: we construct the exchange rates between Germany and the country of assembly of each car model using exchange rate data from OECD.[30]

7.2 Empirical specification

The indirect utility of consumer i, purchasing product j in market t (defined as a state-year pair) is given by:

$$u_{ijt} = \underbrace{x'_{1jt}\beta + \xi^*_{jt}}_{\delta_{jt}} + x'_{2jt}\alpha_i + \varepsilon_{ijt}.$$

The mean utility $\delta_{jt} = x'_{1jt}\beta + \xi^*_{jt}$ captures homogeneous preferences. The variables in x_{1jt} consist of the product characteristics for which we assume that there is no preference heterogeneity and the interaction terms that explain the best the geographical variation observed in Table 7.[31]

The demand shock on product j is decomposed as follows:

$$\xi^*_{jt} = Brand_j + State_t + Year_t + \xi_{jt},$$

[29] https://www.ilo.org/ilostat-files/Documents/Excel/INDICATOR/LAC_4HRL_ECO_CUR_NB_A_EN.xlsx

[30] https://data.oecd.org/conversion/exchange-rates.htm

[31] The choice of the variables that display preference heterogeneity is based on our understanding of the car market and follows current empirical practices for this specific market. However, we understand the limitations of this approach, and we are working on an iterative procedure to select the variables that display consumer heterogeneity.

where $Brand_j$ is a brand fixed effect that captures the unobserved quality of the brand of product j, $State_t$ captures state specific demand shocks that are fixed across time and products and $Year_t$ captures year-specific demand shocks. Therefore, $State_t$ and $Year_t$ play a role in explaining the variation in the overall demand for cars (or equivalently, in the share of the outside option).

The variables in x_{2jt} are the product characteristics that display preference heterogeneity and which we augment with a RC. In our specification, we include the price, the size, and the gasoline dummy in x_{2jt}. We estimate the model assuming different specifications for the distribution of RCs. First, we estimate the model without any consumer heterogeneity. Second, we assume that all the RCs are normally distributed. Finally, we consider a Gaussian mixture on price to increase flexibility with respect to the preferences on price. For each different specification, we perform the specification test developed in section 5 to see how the degree of misspecification evolves as we increase flexibility on the distribution of RCs.

7.3 Estimation

Estimation conditional logit (no heterogeneity) First, we estimate the logit model, and we report the results in Table 8.[32] As expected, we find a negative effect of price and fuel cost on the utility. The interaction terms indicate that the utility derived from size, horsepower, foreign status and gasoline all decrease with income. Moreover, we observe that the aversion to fuel cost decreases over time, which is likely an artifact implied by increasing fuel cost over the years. In contrast, the utility derived from horsepower appears to increase with time. However, these time effects are smaller in comparison with the heterogeneity due to income. To facilitate the interpretation of these results, we consider a household with a €47,000 income in 2018. This corresponds to the mean income in 2018. For this household, the implied effect of size on the utility is negative, whereas a positive utility is derived from higher horsepower, the car's brand being German, height, and gasoline engines.

[32] In Appendix E, we provide results for baseline specifications including the simple conditional logit and the nested logit (with and without state and year fixed effects).

Table 8: Logit estimation

	Baseline		**× Income (/1000)**		**× Pop. density(/100)**		**× Time trend**	
Homogeneous Preferences	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E
Price/income	-2.4	1.3e-01	-	-	-	-	-	-
Fuel Cost	-0.25	8.6e-03	-	-	-	-	0.014	1.7e-03
Size(m^2)	0.15	4.2e-02	-0.0055	8.5e-04	-	-	-	-
Horsepower(KW/100)	2.7	1.8e-01	-0.019	2.4e-03	-	-	-0.081	7e-03
Foreign	0.18	7.1e-02	-0.017	1.4e-03	-	-	-	-
Height(m)	3.5	2.3e-01	-0.0015	4.6e-03	-0.036	4.7e-03	-	-
Gasoline	1.1	6.3e-02	-0.011	1.2e-03	-	-	-	-

Note: Brand, Year and State FE's are included.

Estimation with Gaussian random coefficients We now increase the flexibility in the *traditional* manner, by assuming that the RCs on the price, the size and the gasoline indicator follow a Gaussian distribution. We report the estimates obtained under this new specification in Table 9. The signs for the homogeneous preference parameters in x_{1jt} remain the same and the magnitude of the effects do not change significantly. The sign associated with the mean effect of price remains negative. In contrast, the sign on the mean effects of the size and the gasoline dummy are inverted with respect to the logit specification. This last observation illustrates an important empirical finding: average effects are not invariant to the introduction of preference heterogeneity. In other words, the logit estimates do not necessarily match the means, when we introduce a Gaussian RC. Moreover, the three RCs display high variances and particularly so for the gasoline dummy, which indicate a high level of heterogeneity with respect to these three characteristics.[33]

[33]The estimation is performed using the parametrization proposed in Ketz (2019), which avoids boundary issues at 0 for the variances of the RCs.

Table 9: Traditional BLP (Gaussian RC)

	Baseline		**× Income (/1000)**		**× Pop. density(/100)**		**× Time trend**	
Homogeneous Preferences	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E
Price/income	-	-	-	-	-	-	-	-
Fuel Cost	-0.29	5.1e-03	-	-	-	-	0.031	9.2e-04
Size(m^2)	-	-	-0.0053	3.1e-04	-	-	-	-
Horsepower(KW/100)	0.77	1.5e-02	0.0078	6.8e-04	-	-	-0.12	5.6e-03
Foreign	0.21	5.4e-02	-0.019	1.1e-03	-	-	-	-
Height(m)	3.4	1.1e-02	-0.0088	1.2e-03	-0.032	3.6e-04	-	-
Gasoline	-	-	-0.0028	8.6e-04	-	-	-	-
Gaussian RC	$\hat{\beta}$	S.E	$\hat{\sigma}$	S.E				
Price/income	-2.4	2e-02	0.96	5.9e-03	-	-	-	-
Size(m^2)	-0.37	1.5e-02	0.43	3.6e-03	-	-	-	-
Gasoline	-2.3	4.4e-02	4	4.1e-04	-	-	-	-

Note: Brand, Year and State FE's are included.

Estimation with a Gaussian mixture on the price Finally, we increase the flexibility of the model, by replacing the Gaussian RC on the price variable with a Gaussian mixture of 2 components. We focus on the price as the literature shows that the distribution of price sensitivity is absolutely key for many quantities of interest in IO, including the price elasticities and the pass-through. We report the estimates obtained under this new specification in Table 10. The results point out the presence of two distinct modes in the distribution of the RC associated with price. The two modes reveal the presence of two groups of consumers: the first one with high price sensitivity (with the mean component at -9.6) and the second one with low price sensitivity (with the mean component at -2.5). Moreover, the distribution is heavily asymmetric with the probability of the first mode being 0.9, which entails that the majority of consumers are highly sensitive to price. This last feature is completely absent in the logit and Gaussian specifications, which seem to capture only the first mode of the distribution as we can see in Figure 6. Once again the homogeneous parameters are relatively unchanged with respect to the previous specifications. The Gaussian RC on the gasoline still displays a high variance (the standard deviation of the RC equals 2.8).

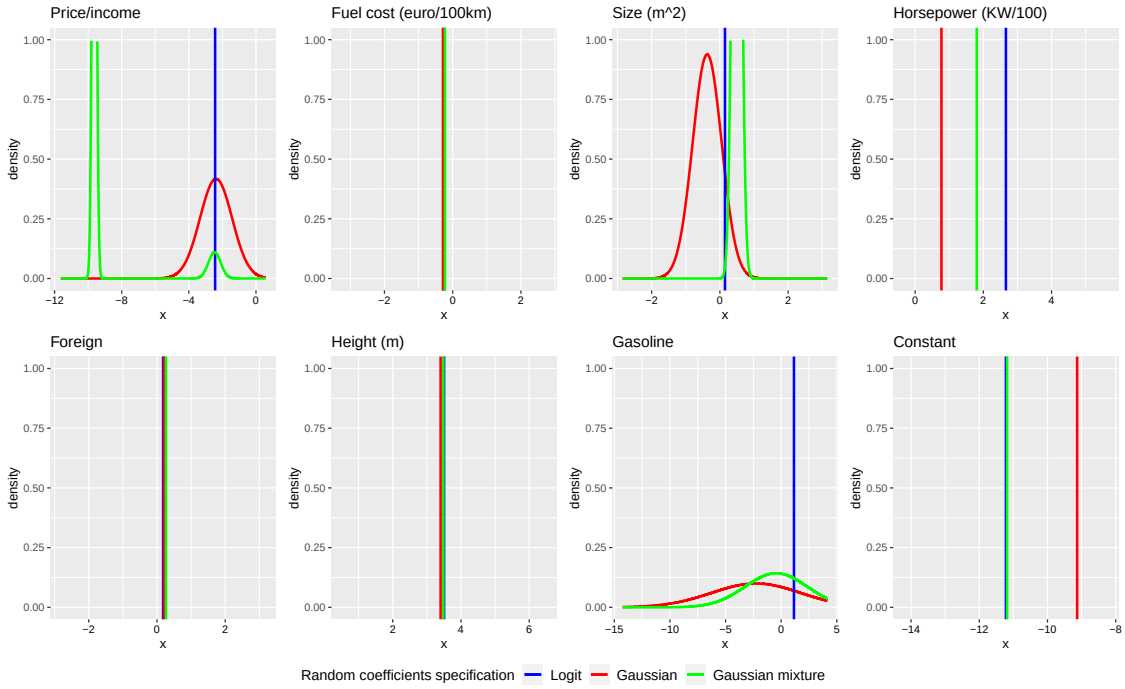
Table 10: Estimation Gaussian mixture on Price

	Baseline		**× Income (/1000)**		**× Pop. density(/100)**		**× Time trend**	
Homogeneous Preferences	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E	$\hat{\beta}$	S.E
Price/income	-	-	-	-	-	-	-	-
Fuel Cost	-0.23	5.8e-03	-	-	-	-	0.026	1e-03
Size(m^2)	-	-	-0.0055	3.7e-04	-	-	-	-
Horsepower(KW/100)	1.8	3.6e-02	-0.0016	1.1e-03	-	-	-0.1	7e-03
Foreign	0.26	6.1e-02	-0.021	1.2e-03	-	-	-	-
Height(m)	3.5	1.1e-02	-0.012	1.2e-03	-0.032	3.7e-04	-	-
Gasoline	-	-	-0.026	1.3e-03	-	-	-	-
Gaussian RC	$\hat{\beta}$	S.E	$\hat{\sigma}$	S.E				
Size(m^2)	0.5	1.9e-02	0.1	6.7e-02	-	-	-	-
Gasoline	-0.45	3.8e-03	2.8	9.1e-03	-	-	-	-
Gaussian Mixture	$\hat{\beta}_1$	S.E	$\hat{\sigma}_1$	S.E	$\hat{\beta}_2$	S.E	$\hat{\sigma}_2$	S.E
Price/income	-9.6	1.8e-02	0.1	1.8e-03	-2.5	1.8e-02	0.35	5.2e-04
Probability	0.9	6.8e-05						

Note: Brand, Year and State FE's are included.

In Figure 6, we plot the estimated distribution of random coefficients under the three specifications we consider. We observe little to no variation in the homogeneous parameters from one specification to the other. The main difference comes from the introduction of the Gaussian mixture on price, which reveals the presence of a large group of highly price sensitive consumers.

Figure 6: Estimated distributions of RCs in the three specifications



Specification test By increasing the flexibility on the distribution of RCs, we recover less precise estimates and the model becomes more difficult to estimate. Thus, it is important to show that the additional flexibility substantially reduces the misspecification of the model. To quantify the degree of misspecification accross the different models, we keep the same set of estimation instruments across the different specifications of RCs and we report the value of the associated Sargan-Hansen J statistics in each case. Moreover, for every model, we follow the procedure developed in section 5 to test if the distribution of RCs on price is well specified. We use the global interval instruments and we denote this test "Interval test". We report the values of the test statistics and the degrees of freedom of the chi-square under the null in Table 11. We observe an important decrease in the Sargan-Hansen J statistic when we transition from the logit to the Gaussian RC. However, the decrease in the Sargan-Hansen J statistic is much larger when we transition from the Gaussian RC on price to the Gaussian mixture, which indicates that the Gaussian mixture performs much better than the simple Gaussian at capturing the underlying heterogeneity in price sensitivity. The interval test displays a similar behavior, with the largest decrease in the test statistic stemming from the transition from the Gaussian RC to the Gaussian mixture.

Table 11: Evolution of misspecification with flexibility

Instruments	Logit			Gaussian RC			Gaussian mixture		
Test	Stat.	Critical val.	DF	Stat.	Critical val.	DF	Stat.	Critical val.	DF
J test	2755.7	40.1	27	2341.7	36.4	24	950.3	33.9	21
Interval test	1331.9	14.1	7	999.4	14.1	7	244.0	14.1	7

7.4 Counterfactual quantities

The objective of this subsection is to illustrate how changes in the distribution of the RC associated with price affect many counterfactual quantities of interest in IO, such as the price elasticities, the marginal costs faced by car manufacturers, and the pass-through of cost. In order to compare our empirical results with the findings in Miravete et al. (2022), we also derive the demand curvature under the different specifications. They show that a large demand curvature is necessary to recover a pass-through larger than one. We now define these different quantities and derive them under the different specifications considered previously. For exposition purposes, we omit the dependence of the market shares in δ_t, x_{2t} and f, and simply write $s_j(\mathbf{p})$ instead of $\rho_j(\delta_t, x_{2t}; f)$, where $\mathbf{p}$ is the price vector. In Appendix E, we provide analytical formulas for every quantity of interest. The quantities of interest are computed using the year 2018, which is the last year of our sample.

- The price elasticity of demand is the ratio of the percentage change in quantity demanded of a product to the percentage change in price. The price elasticity for product j writes as follows: $\eta_j^1(\mathbf{p}) \equiv \frac{p_j}{s_j} \frac{\partial s_j(\mathbf{p})}{\partial p_j}$.

- The demand curvature of the demand function is given by: $\eta_j^2(\mathbf{p}) \equiv \frac{\partial^2 s_j(\mathbf{p})}{\partial p_j^2} \left(\frac{\partial s_j(\mathbf{p})}{\partial p_j} \right)^{-2}$.

- Marginal costs and mark-ups. To recover the marginal costs and the implied mark-ups, we need to make additional assumptions on the supply side. Following the literature, we consider that each multi-product firm $f \in F$ sets prices for its own products in accordance with a Bertrand-Nash equilibrium. The profit of each firm writes:

  $\Pi_f(\mathbf{p}) = \sum_t \sum_{j \in J_f} (p_j - c_j) M_t s_{jt}(\mathbf{p})$,

  where J_f is the set of goods produced by firm f, c_j is the marginal cost for good j, M_t is the market size and $s_j(p)$ is the market share of product j. The first-order condition

with respect to price p_j writes:

$$\sum_t M_t \, s_{jt}(\mathbf{p}) + \sum_t M_t \sum_{j' \in J_f} \left(p_{j'} - c_{j'} \right) \frac{\partial s_{j'}(\mathbf{p})}{\partial p_j} = 0.$$

We gather all the FOCs and rewrite them in matricial form:

$$\mathbf{s}(\mathbf{p}) + (\mathbf{\Delta}(\mathbf{p}))\,(\mathbf{p} - \mathbf{c}) = 0.$$

where $\mathbf{\Delta}(\mathbf{p}) = \sum_t M_t \frac{\partial s_{j'}(\mathbf{p})}{\partial p_j}$ if j' and j are produced by the same firm and equals to zero otherwise. $\mathbf{\Delta}(\mathbf{p})$ is known as the ownership matrix. Assuming that the prices are in equilibrium, one can recover the marginal costs using the following equation:

$$\mathbf{c} = \mathbf{p} - (\Delta(\mathbf{p}))^{-1} \mathbf{s}(\mathbf{p}).$$

The mark-up for product j simply writes: $p_j - c_j$.

- The pass-through of cost is defined as follows. Let us assume that the marginal cost for product j goes from c_j to c'_j (with $c'_j > c_j$), then the cost pass-through equals $\alpha_j = \frac{p'_j - p_j}{c'_j - c_j}$, where p'_j is the new equilibrium price. The pass-through corresponds to the proportion of the cost increase that is transmitted to the price. Following the literature, we derive the pass-through by increasing the marginal costs of each product by 1% and recomputing the marginal cost.

Summary of results We report the median values for the five counterfactual quantities of interest in Table 12. Several remarks are in order. First, the Gaussian mixture yields a much lower price elasticity than the two other specifications. This is related to the emergence of a group of very price sensitive consumers in the mixture specification, which we fail to detect with the logit and Gaussian RC specifications. Moreover, the low price elasticities that we recover in the Gaussian and logit specifications, generate unreasonably low marginal costs (even negative ones as we can see in Figure 7) and excessive mark-ups. In contrast, this problem does not appear with the Gaussian mixture. Finally, to link our results with the findings in Miravete et al. (2022), we now focus on the demand curvature and the pass-through of cost. As expected, the logit displays a curvature and a pass-through equal to 1. In contrast, we can see that the Gaussian mixture displays a larger demand curvature than the other two specifications. This comes from the skewness that the mixture induces in the

distribution of price sensitivity. This last feature implies that the Gaussian mixture yields a pass-through much greater than 1 (1.5 on average). Unfortunately, the negative marginal costs we recover with the Gaussian RC prevent us from computing the pass-through in this case.[34]

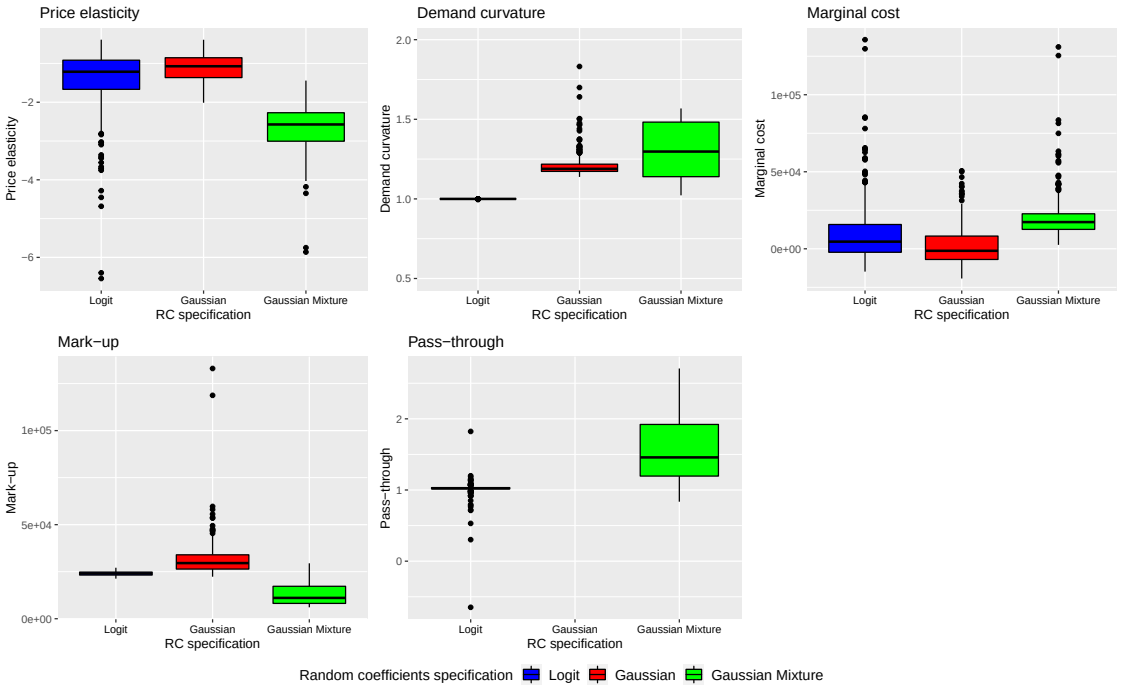
Table 12: Median counterfactual quantities under different specifications on RCs

RC distribution on price	Logit	Gaussian	Gaussian Mixture
Own price-elasticity	-1.2	-1.1	-2.6
Demand curvature	1.0	1.2	1.3
Marginal cost	9,366	1,929	20,105
Mark-up	24,048	29,572	11,066
Pass-through	1.0	-	1.5

In Figure 7, we plot the empirical distributions of the counterfactual quantities. We can see in the plot featuring the distribution of marginal costs that the logit and Gaussian specifications generate negative marginal costs for some of the cars. This is an indication that the price elasticities implied by these specifications are too low in absolute value.

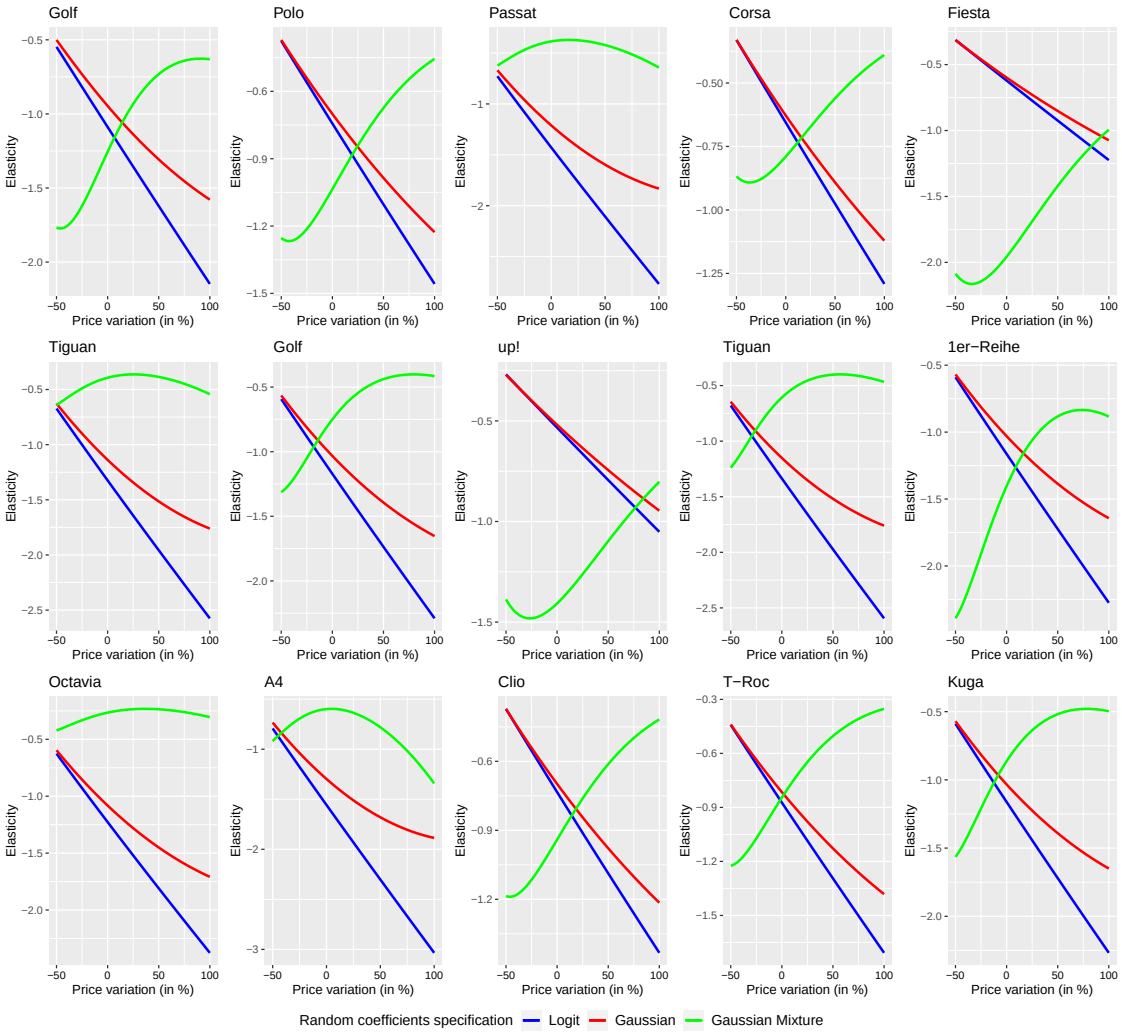
[34]Our algorithm to compute the new equilibrium prices after the change in cost does not converge.

Figure 7: Empirical distribution of counterfactual quantities under different specifications



Finally, in Figure 8, we plot the elasticity functions implied by the different specifications for the 15 most popular cars in our sample. We observe important differences in the elasticities. The Gaussian mixture generates lower price elasticities than the other two specifications. We do the same exercise with the demand curves in Appendix E.

Figure 8: Estimated elasticities under different specifications



8 Conclusion

In this paper, we develop novel econometric tools to parsimoniously increase the flexibility of the distribution of random coefficients in the BLP demand model initiated by Berry et al. (1995). Specifically, we construct novel instruments designed to detect deviations from the true distribution of random coefficients. Building on these instruments, we provide a formal moment-based specification test on the distribution of random coefficients, which allows researchers to test the chosen specification without having to re-estimate the model under a more flexible parametrization. Our instruments are designed to maximize the power of the

test when the distribution of RC is misspecified. By exploiting the duality between estimation and testing, we show that these instruments can also improve the estimation of the BLP model under a flexible parametrization. Our Monte Carlo simulations confirm that the interval instruments we develop in this paper outperform the traditional instruments both for testing and estimating purposes. Finally, we apply these new tools to flexibly estimate the demand for cars in Germany. We show that these tools can be applied to the equally popular mixed logit demand model with individual-level data.

In future works, we plan to see if we can generalize these instruments to other non-linear moment-based models, as well as to the general problem of testing distributional assumptions in structural models. From a broader perspective, our paper is part of an existent discussion on the most effective way to model unobserved preference heterogeneity in structural models. Most empirical frameworks feature a clear trade-off between the degree of flexibility one chooses and the precision of the estimates one obtains. It is thus critical to understand how misspecification on the unobserved heterogeneity affects the counterfactual quantities of interest. In the case of the BLP demand model, our paper and others show that misspecification in the distribution of random coefficients substantially distorts the substitution patterns as well as the shape of the demand curve and, thus, is likely to significantly alter the counterfactual quantities.

Bibliography

Alberini, A. and M. Horvath (2021): "All car taxes are not created equal: evidence from Germany," *Energy Economics*, 100.

Allcott, H. and N. Wozny (2014): "Gasoline prices, fuel economy, and the energy paradox," *Review of Economics and Statistics*, 96, 779–795.

Allen, R. and J. Rehbeck (2020): "Identification of random coefficient latent utility models," *arXiv:2003.00276*.

Amemiya, T. (1974): "The nonlinear two-stage least-squares estimator," *Journal of Econometrics*, 2, 105–110.

Andrews, D. W. and P. Guggenberger (2019): "Identification and singularity-robust inference for moment condition models," *Quantitative Economics*, 10, 1703–1746.

Armstrong, T. B. (2016): "Large market asymptotics for differentiated product demand estimators with economic models of supply," *Econometrica*, 84, 1961–1980.

Bahadur, R. R. (1960): "Stochastic comparison of tests," *The Annals of Mathematical Statistics*, 31, 276–295.

Barahona, N., C. Otero, S. Otero, and J. Kim (2020): "Equilibrium effects of food labeling policies," *SSRN working paper*.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 25, 242–262.

Berry, S. T. and P. A. Haile (2009): "Nonparametric identification of multinomial choice demand models with heterogeneous consumers," *NBER working paper 15276*.

——— (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

Berry, S. T., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T., O. B. Linton, and A. Pakes (2004): "Limit theorems for estimating the parameters of differentiated product demand systems," *The Review of Economic Studies*, 71, 613–654.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

Chetverikov, D. and D. Wilhelm (2017): "Nonparametric instrumental variable estimation under monotonicity," *Econometrica*, 85, 1303–1320.

Compiani, G. (2018): "Nonparametric demand estimation in differentiated products markets," *SSRN working paper*.

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

Crawford, G. S., O. Shcherbakov, and M. Shum (2019): "Quality overprovision in cable television markets," *American Economic Review*, 109, 956–995.

D'Haultfœuille, X., P. Givord, and X. Boutin (2014): "The environmental effect of green taxation: the case of the French *bonus/malus*," *The Economic Journal*, 124, F444–F480.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80, 2231–2267.

Dubois, P., R. Griffith, and M. O'Connell (2018): "The effects of banning advertising in junk food markets," *The Review of Economic Studies*, 85, 396–436.

Dufour, J.-M. and M. L. King (1991): "Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR (1) errors," *Journal of Econometrics*, 47, 115–143.

Dunker, F., S. Hoderlein, and H. Kaido (2022): "Nonparametric identification of random coefficients in endogenous and heterogeneous aggregate demand models," *arXiv preprint: 2201.06140*.

Durrmeyer, I. (2022): "Winners and losers: the distributional effects of the French feebate on the automobile market," *The Economic Journal*, 132, 1414–1448.

Durrmeyer, I. and M. Samano (2018): "To rebate or not to rebate: fuel economy standards vs. feebates," *The Economic Journal*, 128, 3076–3116.

Fox, J. T. and A. Gandhi (2011): "Identifying demand with multidimensional unobservables: a random functions approach," *NBER working paper 17557*.

Fox, J. T., K. il Kim, S. P. Ryan, and P. Bajari (2012): "The random coefficients logit model is identified," *Journal of Econometrics*, 166, 204–212.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185, 162–181.

Gandhi, A. and J.-F. Houde (2019): "Measuring substitution patterns in differentiated-products industries," *NBER working paper 26375*.

Geweke, J. (1981): "The approximate slopes of econometric tests," *Econometrica*, 49, 1427–1442.

Gillingham, K. and S. Houde (2021): "Consumer myopia in vehicle purchases: evidence from a natural experiment," *American Economic Journal: Economic Policy*, 207–238.

Gourieroux, C. and A. Monfort (1995): *Statistics and econometric models*, vol. 2, Cambridge University Press.

Grennan, M. (2013): "Price discrimination and bargaining: empirical evidence from medical devices," *American Economic Review*, 103, 145–177.

Grieco, P. L., C. Murry, and A. Yurukoglu (2022): "The evolution of market power in the US automobile industry," *NBER working paper 29013*.

Grigolon, L., M. Reynaert, and F. Verboven (2018): "Consumer valuation of fuel costs and tax policy: evidence from the European car market," *American Economic Journal: Economic Policy*, 10, 193–225.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Ho, K. and A. Pakes (2014): "Hospital choices, hospital prices, and financial incentives to physicians," *American Economic Review*, 104, 3841–3884.

Huse, C. and N. Koptyug (2022): "Salience and policy instruments: evidence from the auto market," *Journal of the Association of Environmental and Resource Economists*, 9, 345–382.

Ichimura, H. and T. S. Thompson (1998): "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," *Journal of Econometrics*, 86, 269–295.

Ketz, P. (2019): "On asymptotic size distortions in the random coefficients logit model," *Journal of Econometrics*, 212, 413–432.

Kunert, U. (2018): "Diesel fuel and passenger cars receive preferential tax treatment in Europe: reform of taxation needed in Germany," *DIW Weekly Report*, 8, 289–298.

Lee, J. and K. Seo (2015): "A computationally fast estimator for random coefficients logit demand models using aggregate data," *The RAND Journal of Economics*, 46, 86–102.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–177.

Lu, Z., X. Shi, and J. Tao (2021): "Semi-nonparametric estimation of random coefficient logit model for aggregate demand," *SSRN working paper*.

McFadden, D. and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

Miller, N. H., G. Sheu, and M. C. Weinberg (2021): "Oligopolistic price leadership and mergers: the United States beer industry," *American Economic Review*, 111, 3123–3159.

Miller, N. H. and M. C. Weinberg (2017): "Understanding the price effects of the MillerCoors joint venture," *Econometrica*, 85, 1763–1791.

Miravete, E. J., M. J. Moral, and J. Thurk (2018): "Fuel taxation, emissions policy, and competitive advantage in the diffusion of European diesel automobiles," *The RAND Journal of Economics*, 49, 504–540.

Miravete, E. J., K. Seim, and J. Thurk (2022): "Robust pass-through estimation in discrete choice models," *Working paper*.

Nevo, A. (2000): "Mergers with differentiated products: the case of the ready-to-eat cereal industry," *The RAND Journal of Economics*, 31, 395–421.

Newey, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica*, 58, 809–837.

——— (2004): "Efficient semiparametric estimation via moment restrictions," *Econometrica*, 72, 1877–1897.

Petrin, A. (2002): "Quantifying the benefits of new products: the case of the minivan," *Journal of Political Economy*, 110, 705–729.

Reynaert, M. (2021): "Abatement strategies and the cost of environmental regulation: emission standards on the European car market," *The Review of Economic Studies*, 88, 454–488.

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

Roodman, D. (2009): "A Note on the theme of too many instruments," *Oxford Bulletin of Economics and Statistics*, 71, 135–158.

Salanié, B. and F. A. Wolak (2019): "Fast, "robust", and approximately correct: estimating mixed demand systems," *NBER working paper 25726*.

Silvey, S. D. (1959): "The Lagrangian multiplier test," *The Annals of Mathematical Statistics*, 30, 389–407.

Tebaldi, P., A. Torgovitsky, and H. Yang (2019): "Nonparametric estimates of demand in the California health insurance exchange," *NBER working paper 25827*.

Train, K. E. (2008): "EM algorithms for nonparametric estimation of mixing distributions," *Journal of Choice Modelling*, 1, 40–69.

Wang, A. (2022): "Sieve BLP: A semi-nonparametric model of demand for differentiated products," *Journal of Econometrics*.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

A Extension to the mixed logit demand model

The main difference between the BLP demand model and the mixed logit model is that the latter one assumes that the econometrician observes individual data. Let us consider the baseline mixed logit model with no endogeneity and consumer level data.[35] Indirect utility function of consumer i making choice $j \in \{0, 1, \ldots, J\}$ is given by:

$$u_{ij} = x'_{1ij}\beta_0 + x'_{2ij}v_i + \varepsilon_{ij}, \tag{A.12}$$

where

- ε_{ij} is a preference shock that follows a type I extreme value distribution independent of all other variables and across i, j;
- x_{1ij} is a vector of product characteristics interacted with consumer characteristics of size K_1 which display no preference heterogeneity;
- x_{2ij} is a vector of product characteristics interacted with consumer characteristics of size K_2 which display preference heterogeneity;

[35] In the mixed logit case, the absence of endogenous variables here is not an unrealistic assumption as the econometrician can always model unobserved product quality by incorporating product fixed effects into the utility function

- v_i is a vector of random coefficients of size K_2 which jointly follows a joint distribution characterized by a density f;

Each consumer chooses the product that maximizes his or her utility in each market. For any couple $(\tilde{f}, \tilde{\beta})$, demand for product j from consumer i writes:

$$\forall j \neq 0, \quad \rho_j(x_i, \tilde{\beta}, \tilde{f}) = \int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{1ij}\tilde{\beta} + x'_{2ij}v)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\tilde{\beta} + x'_{2ik}v\right\}} \tilde{f}(v)dv.$$

For the outside option, we have:

$$\text{for } j = 0, \quad \rho_j(x_i, \tilde{\beta}, \tilde{f}) = \int_{\mathbb{R}^{K_2}} \frac{1}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\tilde{\beta} + x'_{2ik}v\right\}} \tilde{f}(v)dv.$$

Structural error As we did in the case of the BLP demand model, we can define the structural error generated by $(\tilde{\beta}, \tilde{f})$ as follows. Let y_{ij} equal to 1 if individual i chooses good $j = 0, 1, \ldots, J$.

$$\xi_{ij}(\tilde{\beta}, \tilde{f}) = y_{ij} - \rho_j(x_i, \tilde{\beta}, \tilde{f})$$

By construction, at the true (f, β), we have $\mathbb{E}[\xi_{ij}(\beta, f)|x_i] = \mathbb{E}[y_{ij}|x_i] - \rho_j(x_i, \beta, f) = 0$a.s..

Most powerful instrument and approximations As in the aggregate demand model, let us see how we can construct instruments to detect misspecification in the distribution of RC. Given that the model displays no endogeneity, the set of exogenous variables is simply x_i. We now want to find the transformation of x_i which provides the most detection power against a wrong distribution. With this objective in mind we consider a situation where the econometrician has a candidate (f_0, β_0) and wants to test that the model is well specified, namely: $H_0 : (f, \beta) = (f_0, \beta_0)$. Under an alternative $H_1 : (f, \beta) = (f_a, \beta_a)$, the expression for the Most Powerful Instrument (i.e the instrument which maximizes the correlation between the Structural Error and any instrument in the class of measurable functions of x_i) is the same as previously:

$$\begin{aligned}
\mathbb{E}[\Delta_{0,a}^{\xi_j}|x_{ij}] &= \Delta_j(x_i, f_0, \beta_0, f_a, \beta_a) \\
&= \rho_j(x_i, \beta_0, f_0) - \rho_j(x_i, \beta_a, f_a) \\
&= \int_{\mathbb{R}} \rho_j(x_i, \beta_0, f_0) - \frac{\exp(x'_{1ij}\beta_a + x'_{2ij}v)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_a + x'_{2ik}v\right\}} f_a(v)
\end{aligned}$$

Several remarks are in order. First, contrary to the BLP case, the correction term $\Delta_{0,a}^{\xi_j}$ is a function of the exogenous variables x_i and thus we don't need to estimate its conditional

expectation.Second, β_a and f_a are usually unknown to the econometrician and thus we cannot exploit directly this expression. As did for the BLP case, we propose 2 feasible approximations of the MPI.

- **Global approximation:** we replace the unknown β_a by a known substitute β_0[36]. As for the unknown distribution of RC f_a, we proceed as in the BLP case and we replace the integral with a finite sum. Namely, we have:

$$\mathbb{E}[\Delta_j(x_i, f_0, f_a)|x_i] \approx \sum_{l=1}^{L} \omega_l \underbrace{\left[\rho_j(x_i, \beta_0, f_0) - \frac{\exp(x'_{1ij}\beta_0 + x'_{2ij}v_l)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_0 + x'_{2ik}v_l\right\}}\right]}_{\pi_{j,l}(x_i)}$$

  with $\{v_l\}_{l=1,\dots,L}$ L points chosen in the support of f_a, and ω_l the unknown weights associated with each point

- **Local approximation:** we provide a local approximation which is accurate when f_0 is close to the true density f_a. To derive this local approximation, we need to impose additional restrictions on β_0 and β_a so that $\|\beta_a - \beta_0\| = O\left(\int_{\mathbb{R}^{K_2}} |f_0(v) - f_a(v)|dv\right)$

  Assumption 1 *We assume that $\beta_0 = \beta_0^*$ and $\beta_a = \beta_a^*$ where (β_0^*, β_a^*) are both pseudo true values which maximize the conditional expectation of their respective population log-likelihoods. Namely,*

$$\beta_0^* = \underset{\beta \in \mathbb{R}^{K_1}}{argmax}\ \mathbb{E}\left[L(x_i, y_i, \beta, f_0)\middle|x_i\right] \text{ with } L(x_i, y_i, \beta, f_0) = \sum_{j=0}^{J} \mathbf{1}\{y_{ij} = 1\} \log(\rho_j(x_i, \beta, f_0))$$

$$\beta_a^* = \underset{\beta \in \mathbb{R}^{K_1}}{argmax}\ \mathbb{E}\left[L(x_i, y_i, \beta, f_a)\middle|x_i\right] \text{ with } L(x_i, y_i, \beta, f_a) = \sum_{j=0}^{J} \mathbf{1}\{y_{ij} = 1\} \log(\rho_j(x_i, \beta, f_a))$$

  Now we can derive the following first order approximation of the $\Delta_j(x_i, f_0, \beta_0, f_a, \beta_a)$

  Proposition 1.1
  Under Assumption1, a first order expansion of $\Delta_j(x_i, f_0, \beta_0, f_a, \beta_a)$ around f_0 writes:

$$\begin{aligned}\Delta_j(x_i, f_0, \beta_0, f_a, \beta_a) &= g_j(x_i, \beta_0, f_0) - g_j(x_i, \beta_a, f_a) \\ &= \int_{\mathbb{R}^{K_2}} \frac{exp(x'_{1ij}\beta_0 + x'_{2ij}v)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_0 + x'_{2ik}v\right\}} (f_0(v) - f_a(v))dv + \left.\frac{\partial \rho_j(x_i, \beta, f_a)}{\partial \beta}\right|_{\beta=\beta_0} (\beta_1 - \beta_0) + \mathcal{R}_0\end{aligned}$$

  with $\mathcal{R}_0 = \int_{\mathbb{R}^{K_2}} |f_0(v) - f_a(v)|dv$

[36] in simulations, we find that the homogeneous parameters are usually close to each other even when the distributions are somewhat remote from each other

The proof is in section B. Building on this approximation, we can construct the following local feasible approximation of the MPI:

$$\mathbb{E}[\Delta_j(x_i, f_0, f_a)|x_i] \approx \sum_{l=1}^{L} \bar{\omega}_{1l} \underbrace{\left[\rho_j(x_i, \beta_0, f_0) - \frac{\exp(x'_{1ij}\beta_0 + x'_{2ij}v_l)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_0 + x'_{2ik}v_l\right\}}\right]}_{\bar{\pi}_{1,j,l}(x_i)}$$
$$+ \sum_{l=1}^{L} \bar{\omega}_{2l} \underbrace{\frac{\partial}{\partial \beta}\left\{\frac{exp(x'_{1ij}\beta_0 + x'_{2ij}v_l)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_0 + x'_{2ik}v_l\right\}}\right\}}_{\bar{\pi}_{2,j,l}(x_i)}$$

with $\{v_l\}_{l=1,\dots,L}$ L points chosen in the support of f_a, and $\bar{\omega}_l$ the unknown weights associated with each point. The interval instruments are simply the set $(\bar{\pi}_{1,j,l}(x_i), \bar{\pi}_{2,j,l}(x_i))$.

Specification test

B Proofs

B.1 Identification

In this subsection, we prove that under Assumption A, the distribution of random coefficients f is non-parametrically point identified.

B.1.1 Proof of Proposition 2.1

We want to show that under Assumptions A, the following implication holds:

$$\begin{aligned}(\tilde{f}, \tilde{\beta}) = (f, \beta) \iff & \mathbb{E}[\xi_{jt}(\tilde{f}, \tilde{\beta})|z_{jt}] = 0 \ a.s. \\ \iff & \mathbb{E}\left[\rho_j^{-1}(s_t, x_{2t}, \tilde{f}) - x'_{1jt}\tilde{\beta}\middle| z_{jt}\right] = 0 \ a.s..\end{aligned}$$

Step 1 First, we show that for any random permutation of indexes $j \to j'$, the following equivalence holds:

$$\mathbb{E}[\xi_{jt}|z_{jt}] = 0 \ a.s. \iff \mathbb{E}[\xi_{jt}|z_{j't}] = 0 \ a.s. \ \forall \ j'.$$

First, let us show that the standard exogeneity conditions assumed in Berry and Haile (2014) and in Wang (2022) implies the moment condition we utilize in this paper:

By construction, we can rewrite the exogeneity condition A (i) as follows:

$$\mathbb{E}[\xi_{jt}|z_{jt}] = \sum_{k=1}^{J} \Pr(j=k)\mathbb{E}\left[\xi_{jt}|z_{jt}, j=k\right] = \frac{1}{J}\sum_{k=1}^{J} \mathbb{E}\left[\xi_{jt}|z_{jt}, j=k\right]$$

The exogeneity condition in Wang (2022) assumes: $\forall k, \ \mathbb{E}\left[\xi_{jt}|z_{jt}, j=k\right] = 0$. From what precedes, this condition implies the exogeneity condition $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$ *a.s.* in A (i). This assumption is required for non-parametric identification of the demand functions but not for the non-parametric identification of the distribution of RC.

Now let us prove the identification result. As an artifact for our proof, let us consider a new indexation, which is done exogenously across markets. We denote j' the exogenous indices. Consequently, a same product j doesn't necessarily have the same indices across markets. As the new indexation is done exogenously, we have for any j':

$$\mathbb{E}[\xi_{jt}(\tilde{f},\tilde{\beta})|z_{jt}] = \mathbb{E}[\xi_{jt}(\tilde{f},\tilde{\beta})|z_{jt}, j \to j'] \equiv \underset{j'}{\mathbb{E}}[\xi_{j't}(\tilde{f},\tilde{\beta})|z_{j't}] a.s.$$

$j \to j'$ indicates index j has been changed into j'. Consequently, we have:

$$\mathbb{E}[\xi_{jt}(\tilde{f},\tilde{\beta})|z_{jt}] = 0 \ a.s. \iff \forall j' \ \underset{j'}{\mathbb{E}}[\xi_{j't}(\tilde{f},\tilde{\beta})|z_{j't}] = 0 \ a.s.$$

As a consequence, we can rewrite the initial equivalence as follows:

$$(\tilde{f},\tilde{\beta}) = (f,\beta) \iff \forall j', \ \underset{j'}{\mathbb{E}}[\xi_{j't}(\tilde{f},\tilde{\beta})|z_{j't}] = 0 \ a.s.$$

Given the random permutation $j \to j'$, which is market dependent, we must redefine our matrices and vectors as follows: $\hat{x}_t = M_t x_t$ with $(M_t)_{i,k} = \mathbf{1}\{i = j_t, k = j'_t\}$. Likewise $\hat{s}_t = M_t s_t$. M_t is a random matrix. It is straight forward to show the direct implication.

$$(\tilde{f},\tilde{\beta}) = (f,\beta) \implies \forall j', \quad \underset{j'}{\mathbb{E}}\left[\rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - x'_{1j't}\tilde{\beta}\middle| z_{j't}\right] = \underset{j'}{\mathbb{E}}[\xi_{j't}(f,\beta)|z_{j't}] = 0 \ a.s.$$

The reverse implication is much more intricate to prove and we will exploit other results in the literature. We want to show:

$$(\tilde{f},\tilde{\beta}) \neq (f,\beta) \implies \exists j' \ \middle| \ \underset{j'}{\mathbb{E}}\left[\rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \hat{x}'_{1jt}\tilde{\beta}\middle| z_{j't}\right] = 0 \text{ a.s. does not hold}$$

First, let us assume that $\tilde{f} = f$ and $\tilde{\beta} \neq \beta$, then we have:

$$\rho^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \hat{x}_{1t}\tilde{\beta} = \underbrace{\rho^{-1}(\hat{s}_t, \hat{x}_{2t}, f) - x_{1t}\beta}_{\hat{\xi}_t(f,\beta)} + \hat{x}_{1t}(\beta - \tilde{\beta})$$

By assumption, we have: $P(x'_{1t}x_{1t} \ \ dp) > 0$. M_t is symmetric, idempotent and full rank. As a consequence,

$$P(\hat{x}'_{1t}\hat{x}_{1t} \ \ dp) = P(x'_{1t}M_t x_{1t} \ \ dp) = P(x'_{1t}x_{1t} \ \ dp) > 0$$

Therefore, we have $\forall\ \gamma \neq 0 \in \mathbb{R}^K$,

$$\begin{aligned} P(\gamma'\hat{x}'_{1t}\hat{x}_{1t}\gamma > 0) > P(\hat{x}'_{1t}\hat{x}_{1t} \ \ dp) > 0 &\iff P(\|\hat{x}_{1t}\gamma\|^2 > 0) > 0 \\ &\iff P(\hat{x}_{1t}\gamma \neq 0) > 0 \end{aligned}$$

Thus, $\exists j' \mid x'_{1j't}(\beta - \tilde{\beta}) = 0\ a.s$ does not hold. To conclude, there exists j' such that:

$$\underset{j'}{\mathbb{E}}[\rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, f) - x'_{1j't}\tilde{\beta}|z_{j't}] = \underbrace{\underset{j'}{\mathbb{E}}[\xi_{j't}(f,\beta)|z_{j't}]}_{=0} + \underbrace{\underset{j'}{\mathbb{E}}[x'_{1j't}(\beta - \tilde{\beta})|z_{j't}]}_{=\ 0\ a.s \text{ does not hold from the completeness}}$$

Now let us assume that $\tilde{f} \neq f$ and we want to show that $\forall \tilde{\beta} \in \mathbb{R}^k$, $\exists j'$ such that:

$$\underset{j'}{\mathbb{E}}\left[\rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - x'_{1j't}\tilde{\beta}\middle| z_{jt}\right] = 0\ a.s \ \text{ does not hold}$$

First, note that $\forall j'$,

$$\underset{j'}{\mathbb{E}}\left[\rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - x'_{1j't}\tilde{\beta}|z_{j't}\right] = \underbrace{\underset{j'}{\mathbb{E}}\left[\xi_{j't}(f,\beta)|z_{j't}\right]}_{=0} + \underset{j'}{\mathbb{E}}\left[\rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, f) - x'_{1j't}(\tilde{\beta} - \beta)|z_{j't}\right]$$

Thus, we need to show that $\exists j' \ \left| \ \underset{j'}{\mathbb{E}}\left[\rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, f) - x'_{1j't}(\tilde{\beta} - \beta)\right] = 0\right.$ a.s doesn't hold. From the completeness condition, a sufficient condition is: $\exists j' \left| \ \rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, f) - x'_{1j't}(\tilde{\beta} - \beta) = 0\ a.s\right.$ does not hold. Let $\gamma = (\tilde{\beta} - \beta)$.

By contradiction, it can be easily be shown that $\rho(\hat{\delta}_t, \hat{x}_{2t}, f) - \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, \tilde{f}) \neq 0 \implies \exists j'\ \rho^{-1}_{j'}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) \neq \rho^{-1}_{j}(\hat{s}_t, \hat{x}_{2t}, f) + \gamma' x_{1j't}$. Indeed, assume that $\rho(\hat{\delta}_t, \hat{x}_{2t}, f) - \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, \tilde{f}) \neq$

0 and $\forall j'\ \rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) = \rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, f) + \gamma' x_{1j't}$. Then, we have: $\rho(\rho^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}), \hat{x}_{2t}, \tilde{f}) = \rho(\rho^{-1}(\hat{s}_t, \hat{x}_{2t}, f) + \hat{x}_{1t}\gamma, \hat{x}_{2t}, \tilde{f}) = \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, \tilde{f}) \neq \rho(\hat{\delta}_t, \hat{x}_{2t}, f) = \hat{s}_t$. Therefore, we have a contradiction.

Thus, the next step is to show that $\forall \gamma, \tilde{f} \neq f \implies \rho(\hat{\delta}_t, \hat{x}_{2t}, f_0) - \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, f) = 0$ a.s does not hold.

To this end, we are going to exploit the identification result shown by Wang (2022). Following the notations in this paper, we define $\mu_i = \hat{x}_{1t}\Gamma + \hat{x}_{2t}v_i = \hat{x}_t\mathbf{v}$ with $\mathbf{v}_i = (\Gamma, v_i)$. Here Γ is a degenerate random variable characterized by constant c such that $P(\Gamma = c) = 1$. Let $G_{\mu|\hat{x}_t}$ the distribution of $\mu_i|\hat{x}_t$ under $f^\dagger = (c = 0, f)$ and $G_{\tilde{\mu}|\hat{x}_t}$ the distribution of $\mu_i|\hat{x}_t$ under $\tilde{f}^\dagger = (c = \gamma, \tilde{f})$. The following result is shown in Wang (2022): for any $\hat{\bar{x}}_t \in Supp(\hat{x}_t)$

$$\exists j' \mid \rho_{j'}(\hat{\delta}_t, G_{\mu|\hat{\bar{x}}_t}) - \rho_{j'}(\hat{\delta}_t, G_{\tilde{\mu}|\hat{\bar{x}}_t}) = 0 \text{ on open set } \mathcal{D} \subset \mathbb{R}^J \implies G_{\mu|\hat{\bar{x}}_t} = G_{\tilde{\mu}|\hat{\bar{x}}_t}$$

Note that thanks to the real analytic property of the demand functions ρ, Wang (2022) does not require a full support assumption on $\hat{\delta}_t$

Fix the value of $\hat{x}_t$ as follows: $\hat{x}_t = \bar{M}_t\bar{x}_t = \hat{\bar{x}}_t$.By assumption, there exists $\bar{x}_t \in Supp(x_t)$ such that $\bar{x}_t'\bar{x}_t$ is dp and $\delta_t = \bar{x}_{1t}\beta + \xi_t$ varies on an open set $\bar{\mathcal{D}}$ almost surely.These properties naturally transmit to $\hat{\bar{x}}_t$. The chosen permutation $\bar{M}_t$ doesn't matter. Given the result in Wang (2022), in order to prove that $\rho(\hat{\delta}_t, \hat{x}_{2t}, f_0) - \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, f) = 0$ *a.s* does not hold, we just need to prove that $\forall \gamma,\ \tilde{f} \neq f \implies G_{\tilde{\mu}|\hat{\bar{x}}_t} \neq G_{\mu|\hat{\bar{x}}_t}$. As the density functions are assumed to be continuous, $\tilde{f} \neq f \implies \exists v^* \in \mathbb{R}^{K_2}\ \tilde{F}(v^*) \neq F(v^*)$. Take $x^* = (0_{K_1}, \hat{\bar{x}}_{2t}v^*)' = \hat{\bar{x}}_t(0_{K_1}, v^*)'$:

$$\begin{aligned} &G_{\mu|\hat{\bar{x}}_t}(x^*) = P(x_t\mathbf{v}_i \leq x^*|x_t = \hat{\bar{x}}_t) = P((x_t'x_t)^{-1}x_t'x_t\mathbf{v}_i \leq (x_t'x_t)^{-1}x_t'\bar{x}_t(0_{K_1}, v^*)'|x_t = \hat{\bar{x}}_t) \\ &= (1_{K_1}, P(v_i \leq v^*|x_t = \hat{\bar{x}}_t))' = (1_{K_1}, F(v^*))' \end{aligned}$$

The last equality comes from independence of v_i and x_t. Likewise, $G_{\tilde{\mu}|\hat{\bar{x}}_t}(x^*) = (1\{\gamma > 0\}, \tilde{F}(v^*))'$

Therefore, $\exists x^*,\ \forall \gamma\ \ G_{\tilde{\mu}|\hat{\bar{x}}_t}(x^*) \neq G_{\mu|\hat{\bar{x}}_t}(x^*)$. Following the result in Wang (2022), we have that for all $\gamma \in \mathbb{R}^{K_1}$, $\rho(\hat{\delta}_t, \hat{x}_{2t}, f) - \rho(\hat{\delta}_t + \hat{x}_{1t}\gamma, \hat{x}_{2t}, \tilde{f}) = 0$ *a.s* does not hold which in turn implies that for all $\gamma \in \mathbb{R}^{K_1}$, $\exists j'\ \ \rho_j^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \rho_j^{-1}(\hat{s}_t, \hat{x}_{2t}, f) + \hat{x}_{1jt}'\gamma = 0$ *a.s* does not hold.

To conclude: $\forall \beta \in \mathbb{R}^k$, there exists j' such that:

$$\rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, \tilde{f}) - \rho_{j'}^{-1}(\hat{s}_t, \hat{x}_{2t}, f) - x_{1j't}'(\tilde{\beta} - \beta) = 0 \text{ a.s} \text{ does not hold}$$

which is what we wanted to show.

B.1.2 Proof of Corollary ??

Let us assume that specification $\mathcal{F}_0$, instruments $h_E(z_{jt})$ and weighting matrix yields a unique pseudo true value θ_0.

$$\theta_0 = \underset{\tilde{\theta}}{Argmin}\ \mathbb{E}[\xi_{jt}(f_0(.|\tilde{\lambda},\tilde{\theta})h_E(z_{jt})']W\mathbb{E}[h_E(z_{jt})\xi_{jt}(f_0(.|\tilde{\lambda},\tilde{\theta})]$$

Under $H_0 : f \in \mathcal{F}_0$ and $f = f_0(\cdot|\lambda)$. By the mean independence assumption on the unobserved quality ξ_{jt}, we have at the true $\theta = (\beta, \lambda)$:

$$\xi_{jt}(f_0(.|\lambda),\beta) = \rho_j^{-1}(s_t, x_{2t}, f_0(.|\lambda)) - x_{1jt}'\beta = \xi_{jt} \implies \mathbb{E}[(\xi_{jt}(f_0(.|\lambda),\beta)h_E(z_{jt})] = 0$$

Thus, θ is solution to the previous minimization problem and as the solution is unique: $\theta_0 = \theta$. As a consequence, $\xi_{jt}(f_0(.|\lambda_0),\beta_0)) = \xi_{jt}$ and $\mathbb{E}[\xi_{jt}(f_0(.|\lambda_0),\beta_0)|z_{jt}] = 0$ *as*

Under an alternative specification: $f \notin \mathcal{F}_0$, we know from the identification proof that $\forall\tilde{\theta} = (\tilde{\beta},\tilde{\lambda})$,

$$\mathbb{E}\left[\rho_j^{-1}(s_t, x_{2t}, f_0(.|\tilde{\lambda})) - x_{1jt}'\tilde{\beta}\middle| z_{jt}\right] = 0 \ a.s \ \text{ does not hold}$$

In particular, the last equation holds for $\tilde{\theta} = \theta_0$

B.2 Detecting misspecification: the most powerful instrument

Proof of Proposition 3.1.

- Under $\overline{H}_0 : (f,\beta) = (f_0,\beta_0)$. By assumption, the data are i.i.d. across markets, $\mathbb{E}[\|\xi_{jt}(f_0,\beta_0)h_D(z_{jt})\|^2] = \frac{1}{J}\mathbb{E}[\sum_j\|\xi_{jt}(f_0,\beta_0)h_D(z_{jt})\|^2] < +\infty$, the CLT applies:

$$\frac{1}{\sqrt{TJ}}\sum_{j,t}h_D(z_{jt})\xi_{jt}(f_0,\beta_0) = \frac{1}{\sqrt{TJ}}\sum_{j,t}h_D(z_{jt})\xi_{jt} \underset{T\to+\infty}{\longrightarrow} \mathcal{N}(0,\tilde{\Omega}_0),$$

with:

$$\begin{aligned}
\tilde{\Omega}_0 &= \mathbb{E}\left[\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J} h_D(z_{jt})\xi_{jt}\right)\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J} h_D(z_{jt})\xi_{jt}\right)'\right] \\
&= \frac{1}{J}\mathbb{E}\left[\sum_{j=1}^{J} h_D(z_{jt})h_D(z_{jt})'\xi_{jt}^2 + \sum_{j=1}^{J}\sum_{k\neq j} h_D(z_{jt})h_D(z_{kt})'\xi_{jt}\xi_{kt}\right] \\
&= \frac{1}{J}\mathbb{E}\left[\sum_{j=1}^{J} h_D(z_{jt})h_D(z_{jt})'\xi_{jt}^2\right] + \frac{1}{J}\sum_{j=1}^{J}\sum_{k\neq j}\mathbb{E}\left[h_D(z_{jt})h_D(z_{kt})'\underbrace{\mathbb{E}[\xi_{jt}\xi_{kt}|z_{jt},z_{kt}]}_{=0}\right] \\
&= \mathbb{E}\left[h_D(z_{jt})h_D(z_{jt})'\xi_{jt}^2\right] \\
&= \Omega_0.
\end{aligned}$$

Third line comes from $\xi_{jt} \perp\!\!\!\perp \xi_{kt}|z_t$. By assumption, Ω_0 has a full rank. Thus, we have by the CMT:

$$S_T(h_D, f_0, \beta_0) = TJ\left(\frac{1}{TJ}\sum_{j,t}\xi_{jt}(f_0,\beta_0)h_D(z_{jt})\right)'\hat{\Omega}_0^{-1}\left(\frac{1}{TJ}\sum_{j,t}\xi_{jt}(f_0,\beta_0)h_D(z_{jt})\right) \xrightarrow[T\to+\infty]{d} \chi^2_{|h_D|_0}.$$

- Under H_a' : $\mathbb{E}\left[h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right] \neq 0$. The data are i.i.d. across markets, by the law of large numbers: $\frac{1}{TJ}\sum_{j,t} h_D(z_{jt})\xi_{jt}(f_0,\beta_0) \xrightarrow{P} \mathbb{E}\left[\frac{1}{J}\sum_j h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right]$. It follows by the continuous mapping theorem:

$$\begin{aligned}
\frac{S_T(h_D,f_0,\beta_0)}{T} &\xrightarrow{P} J\mathbb{E}\left[\frac{1}{J}\sum_j h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right]'\Omega_0^{-1}\mathbb{E}\left[\frac{1}{J}\sum_j h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right] \\
&= J\underbrace{\mathbb{E}\left[h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right]'\Omega_0^{-1}\mathbb{E}\left[h_D(z_{jt})\xi_{jt}(f_0,\beta_0)\right]}_{\kappa(h_D,f_0,\beta_0)}
\end{aligned}$$

Under H_a', $\kappa(h_D, f_0, \beta_0)$ is strictly positive because Ω_0 is positive definite. Thence,

$$\begin{aligned}
\forall q \in \mathbb{R},\ \lim_{T\to\infty}\mathbb{P}(S_T(h_D,f_0,\beta_0) > q) &= \lim_{T\to\infty}\mathbb{P}\left(\frac{S(h_D,f_0,\beta_0)-q}{T} > 0\right) \\
&= \mathbb{P}(J\kappa(h_D,f_0,\beta_0) > 0) \\
&= 1,
\end{aligned}$$

where the second equality holds because convergence in probability implies convergence in distribution.

□

Proof of Proposition 3.2. To shorten notations, let $\xi_{jt0} \equiv \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)$, $\xi_{jta} \equiv \xi_{jt}(f_a, \beta_a)$ and ξ_{t0} and ξ_{ta} their stacked versions over j. Likewise, we define $h_D(z_t) = (h_D(z_{1t}), ..., h_D(z_{Jt}))'$ The asymptotic slope of the test writes:

$$\begin{aligned} c_{h_D}(f_a, \beta_a) &= \mathbb{E}\left(\sum_j \xi_{jt0} h_D(z_{jt})\right)' \mathbb{E}\left(\left(\sum_j \xi_{jt0} h_D(z_{jt})\right)\left(\sum_{j'} \xi_{j't0} h_D(z_{j't})\right)'\right)^{-1} \mathbb{E}\left(\sum_j \xi_{jt0} h_D(z_{jt})\right) \\ &= \mathbb{E}(\xi_{t0}' h_D(z_t))\mathbb{E}(h_D(z_t)'\xi_{t0}\xi_{t0}' h_D(z_t))^{-1}\mathbb{E}(h_D(z_t)'\xi_{t0}) \\ &= \mathbb{E}({\Delta_{0,a}^{\xi_t}}' h_D(z_t))\mathbb{E}(h_D(z_t)'\mathbb{E}(\xi_{t0}\xi_{t0}'|z_t)h_D(z_t))^{-1}\mathbb{E}(h_D(z_t)'\Delta_{0,a}^{\xi_t}) \end{aligned}$$

Third line comes from $\mathbb{E}({\Delta_{0,a}^{\xi_t}}' h_D(z_t)) = \mathbb{E}((\xi_{t0} - \xi_{ta})' h_D(z_t)) = \mathbb{E}(\xi_{t0}' h_D(z_t))$ because ξ_{ta} is the true structural error. Then the slope of the test taking $h_D^* = \mathbb{E}(\xi_{t0}\xi_{t0}'|z_t)^{-1}\mathbb{E}(\Delta_{0,a}^{\xi_t}|z_t)$ is equal to:

$$c_{h_D^*}(f_a, \beta_a) = \mathbb{E}\left(\mathbb{E}(\Delta_{0,a}^{\xi_t}|z_t)'\mathbb{E}(\xi_{t0}\xi_{t0}'|z_t)^{-1}\mathbb{E}(\Delta_{0,a}^{\xi_t}|z_t)\right)$$

To finish the proof, we must show that for any set of instruments h_D, we have: $c_{h_D^*}(f_a, \beta_a) \geq c_{h_D}(f_a, \beta_a)$.

Denote $\tilde{h}_D(z_t) = \mathbb{E}(\xi_{t0}\xi_{t0}'|z_t)^{1/2} h_D(z_t)$ and $\tilde{h}_D^*(z_t) = \mathbb{E}(\xi_{t0}\xi_{t0}'|z_t)^{1/2} h_D^*(z_t)$.With these new notations, we have:

$$\begin{aligned} c_{h_D^*}(f_a, \beta_a) - c_{h_D}(f_a, \beta_a) &= \mathbb{E}\left(\tilde{h}_D^*(z_t)'\tilde{h}_D^*(z_t)\right) - \mathbb{E}\left(\tilde{h}_D^*(z_t)'\tilde{h}_D(z_t)\right)\mathbb{E}\left(\tilde{h}_D(z_t)'\tilde{h}_D(z_t)\right)^{-1}\mathbb{E}\left(\tilde{h}_D(z_t)'\tilde{h}_D^*(z_t)\right) \\ &= G'\begin{pmatrix} \mathbb{E}\left(\tilde{h}_D^*(z_t)'\tilde{h}_D^*(z_t)\right) & \mathbb{E}\left(\tilde{h}_D^*(z_t)'\tilde{h}_D(z_t)\right) \\ \mathbb{E}\left(\tilde{h}_D(z_t)'\tilde{h}_D^*(z_t)\right) & \mathbb{E}\left(\tilde{h}_D(z_t)'\tilde{h}_D(z_t)\right) \end{pmatrix} G \\ &= G'\mathbb{E}\left(\tilde{H}\tilde{H}'\right) G \geq 0 \end{aligned}$$

with $\tilde{H} = (\tilde{h}_D^*(z_t), \tilde{h}_D(z_t))'$ and $G = \left(1, -\mathbb{E}\left(\tilde{h}_D^*(z_t)'\tilde{h}_D(z_t)\right)\mathbb{E}\left(\tilde{h}_D(z_t)'\tilde{h}_D(z_t)\right)^{-1}\right)'$

□

Proof of Proposition 3.3.

Under Assumption A, Proposition 2.1 implies the following:

$$\begin{aligned}
\overline{H}_a : (f,\beta) = (f_a,\beta_a) \neq (f_0,\beta_0) \implies & \mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}] \neq 0 \ \ a.s. \\
\implies & \mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]^2 > 0 \ a.s. \\
\implies & \mathbb{E}\left[\mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]^2\right] > 0 \\
\implies & \mathbb{E}\left[\mathbb{E}[\xi_{jt}(f_0,\beta_0)\mathbb{E}[\xi_{jt}(f_0|z_{jt}]|z_{jt}]\right] > 0 \\
\implies & \mathbb{E}\left[\xi_{jt}(f_0,\beta_0)\mathbb{E}[\xi_{jt}(f_0|z_{jt}]\right] > 0 \\
\implies & \overline{H}'_a : \ \mathbb{E}\left[\xi_{jt}(f_0,\beta_0)\underbrace{\mathbb{E}[\Delta_{0,a}^{\xi_{jt}}|z_{jt}]}_{h_D^*(z_{jt})}\right] \neq 0
\end{aligned}$$

Under the same assumptions as 3.1, we have the following:

$$\overline{H}'_a : \ \mathbb{E}\left[\xi_{jt}(f_0,\beta_0)h_D^*(z_{jt})\right] \neq 0 \implies \forall q \in \mathbb{R}^+, \ \ \mathbb{P}(S_T(h_D^*,\mathcal{F}_0,\hat{\theta}) > q) \to 1$$

□

Proof of Proposition 3.4.
Let $\mathcal{H}$ the set of measurable functions of z_{jt}, we want to show under $\bar{H}_a$:

$$\forall \alpha \in \mathbb{R}^*, \ \ \alpha\mathbb{E}[\Delta_{0,a}^{\xi_{jt}}|z_{jt}] \in \arg\max_{h\in\mathcal{H}} \operatorname{corr}(\xi_{jt}(f_0,\beta_0),h(z_{jt}))$$

We proceed in 2 steps. First, we derive the upper bound by showing that for any $h \in \mathcal{H}$, we have:

$$\operatorname{corr}\left(\xi_{jt}(f_0,\beta_0),h(z_{jt})\right) \leq \sqrt{\frac{\operatorname{var}\left(\mathbb{E}[\Delta_{0,a}^{\xi_{jt}}|z_{jt}]\right)}{\operatorname{var}(\xi_{jt}(f_0,\beta_0))}}$$

To do so, we use the definition of the conditional expectation and the Cauchy Schwarz inequality. First, notice that we have: $\mathbb{E}[\Delta_{0,a}^{\xi_{jt}}|z_{jt}] = \mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]$. By definition of the conditional expectation, we have for any $h \in \mathcal{H}$,

$$\mathbb{E}[h(z_{jt})\xi_{jt}(f_0,\beta_0)] = \mathbb{E}[h(z_{jt})\mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]]$$

It follows that:

$$\left|\operatorname{cov}\left(h(z_{jt}),\xi_{jt}(f_0,\beta_0)\right)\right| = \operatorname{cov}\left(h(z_{jt}),\mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]\right) \leq \sqrt{\operatorname{var}(h(z_{jt}))\operatorname{var}\left(\mathbb{E}[\xi_{jt}(f_0,\beta_0)|z_{jt}]\right)}$$

The inequality comes from the Cauchy Schwarz inequality. The result follows by using the definition of the correlation coefficient.

Second, we show that the upper bound is reached by taking for any $\alpha \in \mathbb{R}^*$, $h^*(z_{jt}) = \alpha \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]$.

$$\begin{aligned}\text{cov}\left(\xi_{jt}(f_0, \beta_0), \alpha \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]\right) &= \alpha \text{ cov}\left(\Delta_{0,a}^{\xi_{jt}}, \mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]\right) \\ &= \alpha \text{ var}\left(\mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]\right)\end{aligned}$$

Consequently,

$$\text{corr}\left(\xi_{jt}(f_0, \beta_0), h^*(z_{jt})\right) = \frac{\alpha}{\sqrt{\alpha^2}} \sqrt{\frac{\text{var}\left(\mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]\right)}{\text{var}(\xi_{jt}(f_0, \beta_0))}} \implies \left|\text{corr}\left(\xi_{jt}(f_0, \beta_0), h^*(z_{jt})\right)\right| = \sqrt{\frac{\text{var}\left(\mathbb{E}[\Delta_{0,a}^{\xi_{jt}} | z_{jt}]\right)}{\text{var}(\xi_{jt}(f_0, \beta_0))}}$$

□

B.2.1 Connection with optimal instruments

In the parametric case, the BLP parameter θ is identified by the following non-linear conditional moment restriction $\mathbb{E}[\xi_{jt}(\theta) | z_{jt}] = 0$. The derivation of the optimal instruments in this context has been studied by Amemiya (1974). For an arbitrary choice of $h_E(z_{jt})$, the GMM estimator with the 2-step efficient weighting matrix has the following asymptotic distribution:

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, (\Gamma(\mathcal{F}_0, \theta, h_E)' \Omega(\mathcal{F}_0, h_E)^{-1} \Gamma(\mathcal{F}_0, \theta, h_E))^{-1}\right)$$

with the same notations as previously:

$$\Omega(\mathcal{F}_0, h_E) = \mathbb{E}\left[\left(\sum_j \xi_{jt}(\theta) h_E(z_{jt})\right)\left(\sum_j h_E(z_{jt}) \xi_{jt}(\theta)\right)'\right]$$

$$\Gamma(\mathcal{F}_0, \theta_0, h_E) = \mathbb{E}\left[\sum_j h_E(z_{jt}) \frac{\partial \xi_{jt}(\theta)}{\partial \tilde{\theta}'}\right]$$

For the sake of exposition, we will assume that unobserved demand shock ξ_{jt} is independent across observations, namely: $\mathbb{E}\left[\xi_{jt}(\theta)\xi_{j't}(\theta) | z_t\right] = 0$ for $j \neq j'$. The general case extends naturally. The optimal instrument $h_E^*(z_{jt})$ are chosen to minimize the asymptotic variance covariance matrix. We derive the form of the optimal instruments in the context of BLP by adapting well known results in Chamberlain (1987) and Amemiya (1974)

Lemma 2.1 *Optimal instruments in the BLP model*

In our setting and assuming $f \in \mathcal{F}_0$, the optimal instruments $h_E^*(z_{jt})$ write:

$$h_E^*(z_{jt}) = \mathbb{E}[\xi_{jt}(\theta)^2|z_{jt}]^{-1}\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta)}{\partial \theta}\middle| z_{jt}\right]$$

and the corresponding efficiency bound (obtained by setting $h_E = h_E^*$) writes:

$$V^* = \mathbb{E}\left[\sum_j \mathbb{E}\left[\frac{\partial \xi_{jt}(\theta)}{\partial \theta}\middle| z_{jt}\right]\mathbb{E}\left[\frac{\partial \xi_{jt}(\theta)}{\partial \theta}\middle| z_{jt}\right]' \mathbb{E}[\xi_{jt}(\theta)^2|z_{jt}]^{-1}\right]^{-1}$$

Proof. To shorten the notations, we denote: $\sigma^2(z_{jt}) = \mathbb{E}[\xi_{jt}(\theta)^2|z_{jt}]$ and $d(z_{jt}) = \mathbb{E}\left[\frac{\partial \xi_{jt}(\theta)}{\partial \theta}\middle| z_{jt}\right]$. Likewise, we define

$$\Omega_0(h_E) = \mathbb{E}\left[\sum_j \mathbb{E}[\xi_{jt}(\theta)^2|z_{jt}]h_E(z_{jt})h_E(z_{jt})'\right]$$

We want to prove that for any set of instruments $h_E(z_{jt})$ that $V^*(z_{jt}) - \Gamma_0(h_E)'\Omega_0(h_E)^{-1}\Gamma_0(h_E)$ matrix is semi definite positive.

$$\begin{aligned}
&V^*(z_{jt}) - \Gamma_0(h_E)\Omega_0(h_E)^{-1}\Gamma_0(h_E)' = \\
&= \mathbb{E}\left[\sum_j d(z_{jt})d(z_{jt})'\sigma^2(z_{jt})\right] - \mathbb{E}\left[\sum_j \frac{\partial \xi_{jt}(\theta)}{\partial \theta}h_E(z_{jt})'\right]\Omega_0(h_E)^{-1}\mathbb{E}\left[\sum_j \frac{h_E(z_{jt})\partial \xi_{jt}(\theta)'}{\partial \theta}\right] \\
&= \mathbb{E}\left[\sum_j d(z_{jt})d(z_{jt})'\sigma^{-2}(z_{jt})\right] - \mathbb{E}\left[\sum_j d(z_{jt})h_E(z_{jt})'\right]\mathbb{E}\left[\sum_j \sigma^2(z_{jt})h_E(z_{jt})h_E(z_{jt})'\right]\mathbb{E}\left[\sum_j h_E(z_{jt})d(z_{jt})'\right] \\
&= \mathbb{E}\left[\tilde{\mathbf{D}}(\mathbf{z_{jt}})'\tilde{\mathbf{D}}(\mathbf{z_{jt}})\right] - \mathbb{E}\left[\tilde{\mathbf{D}}(\mathbf{z_{jt}})'\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})\right]\mathbb{E}\left[\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})'\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})'\tilde{\mathbf{D}}(\mathbf{z_{jt}})\right]
\end{aligned}$$

The second line comes from law of iterated expectations. Third line is a matricial way to rewrite the second line. $\tilde{\mathbf{D}}(\mathbf{z_{jt}})$ a matrix which stacks $d(z_{jt})/\sigma(z_{jt})$ over the set of products (each line corresponds to one product j). Likewise, let $\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})$ a matrix which stacks $h_E(z_{jt})\sigma(z_{jt})$ over the set of products (each line corresponds to one product j). Now let us define the following matrices.

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{\mathbf{D}}(\mathbf{z_{jt}}) & \tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}}) \end{pmatrix} \text{ and } \tilde{\mathbf{M}} = \begin{pmatrix} \mathbf{I}_{|\mathbf{z_0}|} & -\mathbb{E}\left[\tilde{\mathbf{D}}(\mathbf{z_{jt}})'\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})\right]\mathbb{E}\left[\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})'\tilde{\mathbf{H}}_\mathbf{E}(\mathbf{z_{jt}})\right]^{-1} \end{pmatrix}'$$

We have: $V^*(z_{jt}) - \Gamma_0(h_E)\Omega_0(h_E)^{-1}\Gamma_0(h_E) = \tilde{\mathbf{M}}'\mathbb{E}[\tilde{\mathbf{X}}'\tilde{\mathbf{X}}]\tilde{\mathbf{M}}$
The matrix above is clearly semi definite positive. □

B.3 Feasible most powerful instrument

B.3.1 Local approximation of the MPI

Proof of Proposition 4.1

Proof. First, we define $s_t^0 = \rho(\delta_t, x_{2t}, f_0(.|\lambda_0))$ with δ_t the true mean utility. From Lemma 2.2 ρ^{-1} is $\mathcal{C}^\infty$ and in particular, ρ^{-1} is $\mathcal{C}^1$. Thus, the Taylor expansion of $\rho^{-1}(s_t^0, x_{2t}, f_0(.|\lambda_0))$ around s_t writes:

$$\rho^{-1}(s_t^0, x_{2t}, f_0(.|\lambda_0)) = \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)) + \frac{\partial \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))}{\partial s}\bigg|_{s=s_t}(s_t^0 - s_t) + o\Big(||s_t^0 - s_t||\Big)$$

$$\delta_t = \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)) + \frac{\partial \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))}{\partial s}\bigg|_{s=s_t}(s_t^0 - s_t) + o\Big(||s_t^0 - s_t||\Big)$$

We now derive an expression for the first derivative of the inverse function. We make use of Lemma 2.3: for any $\delta \in \mathbb{R}^J$, $\frac{\partial \rho(\delta, x_{2t}, f)}{\partial \delta}$ is invertible.

$$\frac{\partial \rho(\rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)), x_{2t}, f_0(.|\lambda_0))}{\partial s} = I_J \iff \frac{\partial \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))}{\partial s}\left(\frac{\partial \rho(\rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)), x_{2t}, f_0(.|\lambda_0))}{\partial \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))}\right) = I_J$$

$$\iff \frac{\partial \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))}{\partial s} = \left(\frac{\partial \rho(\delta_t^0, x_{2t}, f_0(.|\lambda_0))}{\partial \delta}\right)^{-1}$$

with $\delta_t^0 = \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))$.Consequently,

$$\underbrace{\rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)) - \delta_t}_{\Delta(s_t, x_{2t}, f_0, f_a)} = -\left(\frac{\partial \rho(\delta_t^0, x_{2t}, f_0(.|\lambda_0))}{\partial \delta}\right)^{-1}(s_t^0 - s_t) + o\Big(||s_t^0 - s_t||\Big) \tag{B.13}$$

with $\delta_t^0 = \rho_j^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))$

Now let us show that there exists a constant M such that $||s_t^0 - s_t|| \leq M\tau(f_0(.|\lambda_0) - f_a)$. with $\tau(f_0 - f_a) = \int_{\mathbb{R}^{K_2}} |f_0(v|\lambda_0) - f_a(v)|dv$. Norms are equivalent in a finite vectorial space and without loss of generality, we will derive the results with the L_1 norm. By definition:

$$s_t^0 - s_t = \int_{\mathbb{R}^{K_2}} \frac{exp(\delta_t + x_{2t}v)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt} + x'_{2jk}v\right\}} (f_0(v|\lambda_0) - f_a(v))dv$$

Taking the L_1 norm of this vector:

$$\begin{aligned}
||s_t^0 - s_t||_1 &= \sum_{j=1}^{J} \left| \int_{\mathbb{R}^{K_2}} \frac{\exp(\delta_{jt} + x_{2jt}v)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt} + x'_{2jk}v\right\}} (f_0(v|\lambda_0) - f_a(v))dv \right| \\
&\leq \sum_{j=1}^{J} \int_{\mathbb{R}^{K_2}} \underbrace{\left| \frac{\exp(\delta_{jt} + x_{2jt}v)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt} + x'_{2jk}v\right\}} \right|}_{\leq 1} |f_0(v|\lambda_0) - f_a(v)|dv \\
&\leq J \int_{\mathbb{R}^{K_2}} |f_0(v|\lambda_0) - f_a(v)|dv = J\tau(f_0(.|\lambda_0) - f_a)
\end{aligned}$$

This proves the statement. As a consequence, we have: $||s_t^0 - s_t||_1 = O(\tau(f_0(.|\lambda_0) - f_a))$ and $o\big(||s_t^0 - s_t||\big) = o(\tau(f_0(.|\lambda_0) - f_a))$

The problem with the term $s_t^0 - s_t$ is that it is an expression of δ_t which we do not know under misspecification. As we want to be able to compute this approximation of the error term, it is not convenient in practice to have an expression which depends on δ_t. On the other hand, we know δ_t^0 and thus, the simple idea that we exploit is to take a Taylor expansion of the term above around δ_t^0. First, let us remark that from equation B.13, we have that:

$$||\delta_t - \delta_t^0|| = ||\delta_t - \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0)|| = O(||s_t^0 - s_t||) = O(\tau(f_0(.|\lambda_0) - f_a))$$

Now let us take the Taylor expansion of $s_t^0 - s_t$ around δ_t^0:

$$\begin{aligned}
s_t^0 - s_t = &\int_{\mathbb{R}^{K_2}} \frac{exp(\delta_t^0 + x_{2t}v)}{1 + \sum_{k=1}^{J} \exp\left\{\tilde{\delta}_{kt} + x'_{2jk}v\right\}} (f_0(v|\lambda_0) - f_a(v))dv \\
&+ \underbrace{\int_{\mathbb{R}^{K_2}} \frac{\partial}{\partial \delta'} \left\{ \frac{exp(\delta_t^0 + x_{2t}v)}{1 + \sum_{k=1}^{J} \exp\left\{\delta_{kt}^0 + x'_{2jk}v\right\}} \right\} (\delta_t - \delta_t^0)(f_0(v|\lambda_0) - f_a(v))dv + o\Big(||\delta_t - \delta_t^0||\Big)}_{B}
\end{aligned}$$

From what precedes, we know that $o\big(||\delta_t - \delta_t^0||\big) = o(\tau(f_0(.|\lambda_0) - f_a))$. Now, let us show that term B in the previous expansion is also $o(\tau(f_0(.|\lambda_0) - f_a))$. Again taking the L_1 norm:

$$||B||_1 = \sum_{j=1}^{J} \left| \sum_{l=1}^{J} \int_{\mathbb{R}^{K_2}} \frac{\partial}{\partial \delta_l} \left\{ \frac{\exp(\delta_{jt}^0 + x_{2jt}' v)}{1 + \sum_{k=1}^{J} \exp\left\{ \tilde{\delta}_{kt} + x_{2jk}' v \right\}} \right\} (\delta_{lt} - \delta_{lt}^0)(f_0(v|\lambda_0) - f_a(v)) dv \right|$$

$$\leq \sum_{j=1}^{J} \sum_{l=1}^{J} \int_{\mathbb{R}^{K_2}} \underbrace{\left| \frac{\partial}{\partial \delta_l} \left\{ \frac{\exp(\delta_{jt}^0 + x_{2jt}' v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{kt}^0 + x_{2jk}' v \right\}} \right\} \right|}_{\leq 1} |\delta_{lt} - \tilde{\delta}_{lt}||f_0(v|\lambda_0) - f_a(v)| dv$$

$$\leq J^2 \tau(f_0(.|\lambda_0) - f) O(\tau(f_0(.|\lambda_0) - f_a)) = O(\tau(f_0(.|\lambda_0) - f_a)^2) = o(\tau(f_0(.|\lambda_0) - f_a))$$

Thus, $||B||_1 = o(\tau(f_0(.|\lambda_0) - f_a))$ and by combining all the results together, we get the final result. When $f_0(.|\lambda_0)$ gets ''close '' to f_a, we have the following approximation:

$$\Delta(s_t, x_{2t}, f_0, f_a) = \left(\frac{\partial \rho(\delta_t^0, x_{2t}, f_0(.|\lambda_0))}{\partial \delta} \right)^{-1} \int_{\mathbb{R}^{K_2}} \frac{\exp(\delta_t^0 + x_{2t} v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{kt}^0 + x_{2jk}' v \right\}} (f_a(v) - f_0(v|\lambda_0)) dv$$

$$+ o(\tau(f_a - f_0(.|\lambda_0)))$$

$\delta_t^0 = \rho^{-1}(s_t, x_{2t}, f_0(.|\lambda_0))$ and $\tau(f_a - f_0(.|\lambda_0)) = \int_{\mathbb{R}^{K_2}} |f_a(v) - f_0(.|\lambda_0)(v)| dv$. □

B.3.2 Global approximation of the MPI

Derivation of $\Delta_j(s_t, x_{2t}, f_0, f_a)$

Proof.

$$1 = \frac{\rho(\delta_{jt}, x_{2t}, f_a)}{\rho(\delta_{jt}^0, x_{2t}, f_0)} = \frac{\int_{\mathbb{R}^{K_2}} \frac{\exp(\delta_{jt} + x_{2jt}' v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{kt} + x_{2kt}' v \right\}} f_a(v) dv}{\int_{\mathbb{R}^{K_2}} \frac{\exp(\delta_{jt}^0 + x_{2jt}' v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{kt}^0 + x_{2kt}' v \right\}} f_0(v) dv} \iff \frac{\exp(\delta_{jt}^0)}{\exp(\delta_{jt})} = \frac{\int_{\mathbb{R}^{K_2}} \frac{\exp(x_{2t} v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{kt} + x_{2kt}' v \right\}} f_a(v) dv}{\int_{\mathbb{R}^{K_2}} \frac{\exp(x_{2jt}' v)}{1 + \sum_{k=1}^{J} \exp\left\{ \delta_{jt}^0 + x_{2kt}' v \right\}} f_0(v) dv}$$

□

B.3.3 Approximation of the MPI in the mixed logit case

Proof of Proposition 1.1. By definition, we have:

$$g_j(x_i, ., f) : \mathbb{R}^{K_1} \to [0, 1]$$

$$\beta \mapsto \int_{\mathbb{R}^{K_2}} \frac{\exp\left\{ x_{ij1}' \beta + x_{2ij}' v \right\}}{1 + \sum_{k=1}^{J} \exp\left\{ x_{ik1}' \beta + x_{2ik}' v \right\}} f(v) dv$$

g is $\mathcal{C}^\infty$ on $\mathbb{R}^{K_1}$. Thus, we can take a first order Taylor expansion of $g_j(x_i, ., f_1)$ around β_0^*:

$$g_j(x_i, \beta_1, f_1) = g_j(x_i, \beta_0^*, f_1) + \frac{\partial g(x_i, \beta, f_1)}{\partial \beta}\bigg|_{\beta=\beta_0^*} (\beta_1 - \beta_0^*) + o(||\beta_1 - \beta_0^*||)$$

This yields immediately,

$$g(x_i, \beta_0^*, f_0^*) - g(x_i, \beta_1, f_1) = \int_{\mathbb{R}^{K_2}} \frac{exp(x'_{1ij}\beta_0^* + x'_{2ij}v)}{1 + \sum_{k=1}^{J} \exp\left\{x'_{1ik}\beta_0^* + x'_{2ik}v\right\}} (f_0^*(v) - f_1(v))dv +$$
$$\frac{\partial g(x_i, \beta, f_1)}{\partial \beta}\bigg|_{\beta=\beta_0} (\beta_a - \beta_0) + o(||\beta_- \beta_0^*||)$$

Now let us show that $||\beta_1 - \beta_0^*|| = ...$

By construction, the pseudo true values β_0^* and β_1^* maximize the conditional expectation of the log-likelihood:

$$\beta_0^* = \underset{\beta \in \mathbb{R}^{K_1}}{argmax}\ \mathbb{E}\left[L(x_i, y_i, \beta, f_0^*)\middle| x_i\right] \text{ with } L(x_i, y_i, \beta, f_0^*) = \sum_{j=0}^{J} \mathbf{1}\{y_{ij} = 1\} log(g_j(x_i, \beta, f_0^*))$$

The same goes for β_1^*:

$$\beta_1^* = \underset{\beta \in \mathbb{R}^{K_1}}{argmax}\ \mathbb{E}\left[L(x_i, y_i, \beta, f_1^*)\middle| x_i\right] \text{ with } L(x_i, y_i, \beta, f_1^*) = \sum_{j=0}^{J} \mathbf{1}\{y_{ij} = 1\} log(g_j(x_i, \beta, f_1^*))$$

When the true distribution of RC is f_1, we have:

$$\mathbb{E}\left[L(x_i, y_i, \beta, f_0^*)\middle| x_i\right] = \sum_{j=0}^{J} g_j(x_i, \beta_1^*, f_1) log(g_j(x_i, \beta, f_0^*))$$
$$\mathbb{E}\left[L(x_i, y_i, \beta, f_1)\middle| x_i\right] = \sum_{j=0}^{J} g_j(x_i, \beta_1^*, f_1) log(g_j(x_i, \beta, f_1))$$

□

B.4 Specification test: composite hypothesis

In this section, we prove Theorem 5.1, which is the main asymptotic result of the paper. The section is organized as follows. First, we establish the equivalence between the moment

condition around which we build our test $\mathbb{E}\left[\sum_{jt}\xi_{jt}(f_0(\cdot|\lambda_0),\beta_0)h_D(z_{jt})\right] = 0$ and the one characterizing $H_0' : \mathbb{E}\left[\xi_{jt}(f_0(\cdot|\lambda_0),\beta_0)h_D(z_{jt})\right] = 0$. Then, we introduce the notations used in the proofs and we decompose $\widehat{\xi}$ according to the BLP approximations. Second, we provide technical lemmas which prove that under the assumptions in E, the BLP approximations vanish asymptotically. Third, we prove that the BLP estimator is consistent and asymptotically normal. Finally, we prove the main theorem and we show that under the null the test is pivotal in the 2 polar cases described in the main text.

B.4.1 Equivalence between moment conditions

Let $h_D(z_{jt})$ our detection instruments. For conciseness, we omit the dependence in f_0 and denote $\xi_{jt}(f_0(\cdot|\lambda_0),\beta_0) = \xi_{jt}(\theta_0)$. We want to prove that the following two moment conditions are equivalent:

$$\mathbb{E}\left[\xi_{jt}(\theta_0)h_D(z_{jt})\right] = 0 \iff \mathbb{E}\left[\sum_{j=1}^{J}\xi_{jt}(\theta_0)h_D(z_{jt})\right] = 0$$

Let R_t a categorial random variable which exogenously selects a product j with probability $\frac{1}{J}$. Formally, we have $(\xi_{jt}(\theta_0), z_{jt}) \perp R_{jt}$. By construction, we have:

$$\begin{aligned}\mathbb{E}\left[\xi_{jt}(\theta_0)h_D(z_{jt})\right] &= \sum_{k=1}^{J}\mathbb{E}\left[\xi_{kt}(\theta_0)h_D(z_{kt})R_{kt}\right] = \sum_{k=1}^{J}\mathbb{E}\left[\xi_{kt}(\theta_0)h_D(z_{kt})\right]\mathbb{E}[R_{kt}] \\ &= \frac{1}{J}\mathbb{E}\left[\sum_{k=1}^{J}\xi_{kt}(\theta_0)h_D(z_{kt})\right]\end{aligned}$$

The second line results from independence of $(\xi_{jt}(\theta_0), z_{jt})$ and R_{jt}. This proves the result.

B.4.2 Notations

In the proofs, we will adopt the following notations. If the derivations are done under the parametric assumption $H_0 : f \in \mathcal{F}_0$ then we omit the dependence in f_0 and interchangeably use $\xi_{jt}(f_0(.|\lambda),\beta)$ and $\xi_{jt}(\theta)$. We also omit the dependence of the BLP pseudo true value in W and $h_E(z_{jt})$ [37]. Then define the following objectives of the GMM minimization

[37] The BLP pseudo true value depends on W and $h_E(z_{jt})$ when the model is misspecified

$$\hat{\mathcal{Q}}_T(\tilde{\theta}) = \left(\frac{1}{T}\sum_{j,t}\widehat{\xi}_{jt}(\tilde{\theta})h_E(z_{jt})\right)' \hat{W}\left(\frac{1}{T}\sum_{j,t}\widehat{\xi}_{jt}(\tilde{\theta})h_E(z_{jt})\right)$$

$$\mathcal{Q}_T(\tilde{\theta}) = \left(\frac{1}{T}\sum_{j,t}\xi_{jt}(\tilde{\theta})h_E(z_{jt})\right)' \hat{W}\left(\frac{1}{T}\sum_{j,t}\xi_{jt}(\tilde{\theta})h_E(z_{jt})\right)$$

$$\mathcal{Q}(\tilde{\theta}) = \mathbb{E}\left[\sum_{j}\xi_{jt}(\tilde{\theta})h_E(z_{jt})\right]' W \mathbb{E}\left[\sum_{j}\xi_{jt}(\tilde{\theta})h_E(z_{jt})\right]$$

We also define the following moments

$$\hat{g}_T(\tilde{\theta}, h) = \frac{1}{T}\sum_{jt}\hat{\xi}_{jt}(\tilde{\theta})h(z_{jt})$$

$$g_T(\tilde{\theta}, h) = \frac{1}{T}\sum_{jt}\xi_{jt}(\tilde{\theta})h(z_{jt})$$

$$g(\tilde{\theta}, h) = \mathbb{E}\left[\sum_{j}\xi_{jt}(\tilde{\theta})h(z_{jt})\right]$$

And recall the definition of $\Gamma(\mathcal{F}_0, \tilde{\theta}, h)$ which is used interchangeably with $\Gamma(\tilde{\theta}, h)$

$$\hat{\Gamma}_T(\tilde{\theta}, h) = \frac{1}{T}\sum_{j,t}h(z_{jt})\frac{\partial}{\partial\theta}\widehat{\xi}_{jt}(\tilde{\theta})'$$

$$\Gamma_T(\tilde{\theta}, h) = \frac{1}{T}\sum_{j,t}h(z_{jt})\frac{\partial}{\partial\theta}\xi_{jt}(\tilde{\theta})'$$

$$\Gamma(\tilde{\theta}, h) = \mathbb{E}\left[\sum_{j}h(z_{jt})\frac{\partial}{\partial\theta}\xi_{jt}(\tilde{\theta})'\right]$$

Furthermore, unless specified, all limits are taken with respect to T; Additionally, we denote by the expression $X = o_P(T^\kappa)$ a random variable or statistic which is asymptotically degenerate of order T^a, ie $X = o_P(T^\kappa) \Leftrightarrow \forall e > 0\ \mathbb{P}(|X|T^{-\kappa} > e) \underset{T\to\infty}{\to} 0$, and denote by $X = O_p(T^\kappa)$ a random variable which is (bounded in probability) of order T^κ, ie $\forall e_1 > 0 \exists e_2 > 0, \exists T_N : \forall T \geqslant T_N\ \mathbb{P}(|X|T^{-\kappa} > e_2) < e_1$. Properties of $o_P(1)$ and $O_P(1)$ random variables are used throughout these proofs.

B.4.3 Feasible Structural Error and BLP approximations

We now decompose the difference between the true structural error $\xi_{jt}(\tilde{\theta})$ and the feasible structural error $\hat{\xi}_{jt}(\tilde{\theta})$ in terms of the different approximations involved in the derivation of the feasible structural error $\hat{\xi}_{jt}(\tilde{\theta})$. In market t given an assumption $\mathcal{F}_0$, a parameter $\tilde{\lambda}$, market shares s_t and product characteristics with preference heterogeneity x_{2t} there exists a unique $\delta_t \in \mathbb{R}^J$ such that $s_t = \rho(\delta_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ (Brouwer's fixed point theorem, see Berry (1994)) so that $\delta_t = \rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$. There is no closed form for $\rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ so the NFP algorithm is used. Denote as C the contraction used to find the mean utilities which solve the demand equal market share constraint

$$C(\cdot, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) : \delta \in \mathbb{R}^J \mapsto \delta + log(s_t) - log(\rho(\delta, x_{2t}, f_0(\cdot|\tilde{\lambda})))$$

So that for some starting mean utility $\delta_0 \in \mathcal{B} \subset \mathbb{R}^J$ where $\mathcal{B}$ is bounded, the mean utility obtained via NFP at the limit is equal to the unique vector which solves the constraint

$$\delta_t(f_0(\cdot|\tilde{\lambda})) = \rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) = \lim_{H\to\infty} C^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$$

Similarly the error generated by $(f_0(\cdot|\tilde{\lambda}, \tilde{\beta}))$ can be obtained from NFP at the limit

$$\xi_t(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) = \delta_t(f_0(\cdot|\tilde{\lambda})) - x_{1t}\tilde{\beta} = \lim_{H\to\infty} C^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) - x_{1t}\tilde{\beta}$$

This way we obtain a vector of mean utilities for each market t. There are 3 approximations to consider, market shares are not truly observed, the demand integral has to be simulated, and the contraction is never taken to its limit, so define $\widehat{\xi}(f_0, \tilde{\lambda}))$ $\widehat{\delta}(f_0, \tilde{\lambda}))$ and $\hat{C}$ for some starting value δ_0

$$\widehat{\xi}_t(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) = \hat{C}^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) - x_{1jt}\tilde{\beta}, \qquad \widehat{\delta}(f_0, \tilde{\lambda})) = \hat{C}^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0, \tilde{\lambda}))$$
$$\hat{C} : \delta \mapsto \delta + log(\hat{s}_t) - log(\hat{\rho}(\delta, x_{2t}, f_0(\cdot|\lambda_0)))$$

Consequently, we decompose the difference between the error generated by $(f_0(\cdot|\tilde{\lambda}), \tilde{\beta})$ and its feasible approximation into 3 differences

$$\begin{aligned}
\xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) - \widehat{\xi}_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) &= \delta_{jt}(f_0(\cdot|\tilde{\lambda})) - \widehat{\delta}_{jt}(f_0(\cdot|\tilde{\lambda})) \\
&= \lim_{H\to\infty} C_j^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))) - \hat{C}_j^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) \\
&= \lim_{H\to\infty} C_j^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))) - C_j^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) \\
&\quad + C_j^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))) - C_j^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) \\
&\quad + C_j^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) - \hat{C}_j^{(H)}(\delta_0, \hat{s}_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) \\
&\equiv \rho_j^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})) - D_j(\rho, s_t, \tilde{\lambda}) \\
&\quad + D_j(\rho, s_t, \tilde{\lambda}) - D_j(\rho, \hat{s}_t, \tilde{\theta}) \\
&\quad + D_j(\rho, \hat{s}_t, \tilde{\theta}) - D_j(\hat{\rho}, \hat{s}_t, \tilde{\theta})
\end{aligned}$$

In the fourth line, we simply introduce shortened notations for the same objects.

B.4.4 Technical lemmas

The 1st and 2nd lemma establish the smoothness of ρ^{-1} and the invertibility of the Jacobian matrix of ρ with respect to δ. In the 3rd lemma, we derive the Lipschitz constant of the contraction and we prove that it is bounded away from 0 and 1. The 4th lemma ensures that for key moments and quantities the BLP approximations can be ignored uniformly asymptotically.

Lemma 2.2 *ρ^{-1} is $\mathcal{C}^\infty$*

Proof. We know that the demand function ρ is $\mathcal{C}^\infty$ and invertible on $\mathbb{R}^J$. Moreover,$\forall \delta \in \mathbb{R}^J$, $\frac{\partial \rho(\delta, x_{2t}, f)}{\partial \delta} \neq 0$. As a consequence, $\rho^{-1} : [0,1]^J \to \mathbb{R}^J$ the inverse demand function is also $\mathcal{C}^\infty$. □

Lemma 2.3 *For any $\delta \in \mathbb{R}^J$, $\frac{\partial \rho(\delta, x_{2t}, f)}{\partial \delta}$ is invertible.*

Proof. $\frac{\partial \rho}{\partial \delta}$ is a $J \times J$ matrix such that $\left(\frac{\partial \rho}{\partial \delta}\right)_{j,k}$ is:

$$\frac{\partial \rho_j(\delta_t, x_{2t}, f)}{\partial \delta_{kt}} = \begin{cases} \int \mathcal{T}_{jt}(v)\left(1 - \mathcal{T}_{kt}(v)\right) f(v) dv & if \quad j = k \\ -\int \mathcal{T}_{jt}(v)\mathcal{T}_{kt}(v) f(v) dv & if \quad j \neq k \end{cases}$$

with $\mathcal{T}_{jt}(v) \equiv \frac{exp\{\delta_{jt} + x'_{2jt} v\}}{1 + \sum_{j'=1}^{J} exp\{\delta_{j't} + x'_{2j't} v\}}$

One can easily check that $\frac{\partial \rho}{\partial \delta}$ is strictly diagonally dominant. Indeed for each row j:

$$\left|\frac{\partial \rho_j(\delta_t, x_{2t}, f)}{\partial \delta_{kt}}\right| - \sum_{k \neq j} \left|\frac{\partial \rho_j(\delta_t, x_{2t}, f)}{\partial \delta_{kt}}\right| = \int \mathcal{T}_{jt}(v) \underbrace{\left(1 - \sum_{k=1}^{J} \mathcal{T}_{kt}(v)\right)}_{>0} f(v) dv > 0$$

□

Lemma 2.4 (Contraction mapping Lipschitz constant)
Given parametric assumption $\mathcal{F}_0$, under assumptions B-E, assume that starting mean utility δ_0 is in $\mathcal{B}$ where $\mathcal{B}$ is compact, then without loss of generality there exists some $(\underline{a}, \bar{a}) \in \mathbb{R}^2$ with $\bar{a} > \underline{a}$ such

that for any $b \in \mathcal{B}$ for any $j = 1, \ldots, J$ $\underline{a} \leqslant b_j \leqslant \bar{a}$, furthermore denote by $\mathcal{X}$ the compact support of x_{2jt}. Then on $\mathcal{B}$ the map $C(\cdot, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}_0))$ is a contraction with Lipschitz constant

$$\epsilon = \max_{j=1,\ldots,J} \sup_{a \in \mathcal{B}, b \in [0;\bar{a}-\underline{a}]^J, x_2 \in \mathcal{X}, \tilde{\lambda} \in \Lambda_0} 1 - \frac{\int \frac{exp(a_j+b_j+x'_{2j}v)}{\left(1+\sum_k \exp(a_k+b_k+x'_{2k}v)\right)^2} f_0(v|\tilde{\lambda})dv}{\int \frac{exp(a_j+b_j+x'_{2j}v)}{1+\sum_k \exp(a_k+b_k+x'_{2k}v)} f_0(v|\tilde{\lambda})dv}$$

which is in $(0;1)$

Proof. This proof is inspired by the proof of the Theorem in Appendix 1 of Berry et al. (1995). Let $C_j(\cdot) \equiv C(\cdot, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}_0))$, we first determine the partial derivative of $C_j(\cdot)$

$$\begin{aligned}
\frac{\partial C_j(a)}{\partial a_j} &= 1 - \frac{1}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} \int \frac{exp(a_j + x'_{2kt}v)(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v)) - exp(2(a_j + x'_{2kt}v))}{(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv \\
&= \frac{1}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} \int \frac{exp(2(a_j + x'_{2jt}v))}{(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv \\
\frac{\partial C_j(a)}{\partial a_{j'}} &= \frac{1}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} \int \frac{exp(a_j + x'_{2jt}v)exp(a_{j'} + x'_{2j't}v)}{(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv
\end{aligned}$$

Note that for any $j = 1, \ldots, J$ all partial derivatives of $C_j(\cdot)$ are strictly positive and that the sum of its derivatives evaluated in a equals

$$\begin{aligned}
\sum_{k=1}^J \frac{\partial C_j(a)}{\partial a_k} &= \frac{1}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} \int \frac{exp(a_j + x'_{2jt}v) \sum_{k=1}^J exp(a_k + x'_{2kt}v)}{(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv \\
&= \frac{1}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} \int \frac{exp(a_j + x'_{2jt}v)(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v) - 1)}{(1 + \sum_{k=1}^J exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv \\
&= 1 - \frac{\int \frac{exp(a_j+x'_{2jt}v)}{(1+\sum_{k=1}^J exp(a_k+x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))}
\end{aligned}$$

For any $(a_1, a_2) \in \mathcal{B}^2$ let $\tilde{a} = (||a_1 - a_2||_\infty, \ \dots \ , ||a_1 - a_2||_\infty) \in \mathbb{R}^J$ then

$$\begin{aligned} C_j(a_1) - C_j(a_2) &= C_j(a_2 + a_1 - a_2) - C_j(a_2) \leqslant C_j(a_2 + \tilde{a}) - C_j(a_2) \\ &\leqslant \int_{0^J}^{||a_1 - a_2||_\infty^J} \frac{\partial C_j(a_2 + b)}{\partial a} db \\ &\leqslant ||a_1 - a_2||_\infty \sup_{a \in \mathcal{B}, b \in [0; \bar{a} - \underline{a}]^J} \sum_{k=1}^{J} \frac{\partial C_j(a + b)}{\partial a_k} \\ &\leqslant ||a_1 - a_2||_2 \max_{j=1,..J} \sup_{a \in \mathcal{B}, b \in [0; \bar{a} - \underline{a}]^J, x_2 \in \mathcal{X}, \tilde{\lambda} \in \Lambda_0} \sum_{k=1}^{J} \frac{\partial C_j(a + b)}{\partial a_k} \\ &\equiv ||a_1 - a_2||_2 \ \epsilon \end{aligned}$$

where the 1st inequality holds because $C_j(\cdot)$ is increasing in all its inputs, the 2nd inequality holds by the fundamental theorem of calculus and by the total derivative formula, the 3rd and 4th inequalities hold by properties of norms.

We now prove that $\sup_{a \in \mathcal{B}, b \in [0; \bar{a} - \underline{a}]^J, \tilde{\lambda} \in \Lambda_0} \sum_{k=1}^{J} \frac{\partial C_j(a+b)}{\partial a_k} \in (0; 1)$ which will imply that $\epsilon \in (0; 1)$. To do so we have to prove that $\sum_{k=1}^{J} \frac{\partial C_j(a, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial a_k}$ is continuous in $(a, x_{2t}, \tilde{\lambda})$ and takes values in $(0; 1)$ almost surely, this way because $\mathcal{B}$, $\mathcal{X}$ and Λ_0 are compact by Weierstrass' extreme value Theorem the sum of partial derivatives will also take values in a compact which is inside $(0; 1)$, then the supremum will become a maximum which can be attained and which is inside $(0; 1)$. The sum of partial derivatives is almost surely in $(0; 1)$ because

$$
\begin{aligned}
&\int \frac{exp(a_j + x'_{2jt}v)}{(1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv - \rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda})) \\
&= \int \frac{exp(a_j + x'_{2jt}v)}{(1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv - \int \frac{exp(a_j + x'_{2jt}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)} f_0(v|\tilde{\lambda})dv \\
&= -\int \frac{exp(a_j + x'_{2jt}v)\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)}{(1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv < 0 \\
&\Rightarrow \frac{\int \frac{exp(a_j + x'_{2jt}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)} f_0(v|\tilde{\lambda})dv}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} < 1 \\
&\Rightarrow \sum_{k=1}^{J} \frac{\partial C_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial a_k} = 1 - \frac{\int \frac{exp(a_j + x'_{2jt}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)} f_0(v|\tilde{\lambda})dv}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} > 0 \\
&\quad - \frac{\int \frac{exp(a_j + x'_{2jt}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)} f_0(v|\tilde{\lambda})dv}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} < 0 \\
&\Rightarrow \sum_{k=1}^{J} \frac{\partial C_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial a_k} = 1 - \frac{\int \frac{exp(a_j + x'_{2jt}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v)} f_0(v|\tilde{\lambda})dv}{\rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))} < 1
\end{aligned}
$$

Continuity of the sum of the partial derivatives in (a, x_{2t}) is trivial, continuity in $\tilde{\lambda}$ also holds because $f_0(\cdot|\tilde{\lambda})$ must be continuously differentiable via Assumption D. $\forall e_1 > 0, \exists e_2 : \forall(\lambda_1, \lambda_2) : ||\lambda_1 - \lambda_2||_2 \leqslant e_2$ implies $|f_0(v|\lambda_1) - f_0(v|\lambda_2)| < e_1$ for all v which in turn implies

$$
\begin{aligned}
\forall x_2 \in \mathcal{X}, \forall a \in \mathcal{B} \quad & \left| \int \frac{exp(a_j + x'_{2j}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2k}v)} (f_0(v|\lambda_1) - f_0(v|\lambda_2))dv \right| \\
& \leqslant \int \frac{exp(a_j + x'_{2j}v)}{1+\sum_{k=1}^{J} exp(a_k + x'_{2k}v)} |f_0(v|\lambda_1) - f_0(v|\lambda_2)|dv \leqslant e_1 \\
& \left| \int \frac{exp(a_j + x'_{2j}v)}{(1+\sum_{k=1}^{J} exp(a_k + x'_{2k}v))^2} (f_0(v|\lambda_1) - f_0(v|\lambda_2))dv \right| \leqslant e_1
\end{aligned}
$$

thus both $\tilde{\lambda} \mapsto \rho_j(a, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ and $\tilde{\lambda} \mapsto \int \frac{exp(a_j + x'_{2jt}v}{(1+\sum_{k=1}^{J} exp(a_k + x'_{2kt}v))^2} f_0(v|\tilde{\lambda})dv$ are continuous and so is their ratio. □

Lemma 2.5 (Uniform convergence of objective function wrt BLP approximations)

Given parametric assumption $\mathcal{F}_0$, under assumptions B-E and $\forall h$ which satisfies D

$$\sup_{\tilde{\theta}\in\Theta_0} \sqrt{T}||\hat{g}_T(\tilde{\theta},h) - g_T(\tilde{\theta},h)||_2 \xrightarrow{P} 0$$
$$\sup_{\tilde{\theta}\in\Theta_0} ||\hat{\Gamma}_T(\tilde{\theta},h) - \Gamma_T(\tilde{\theta},h)||_2 \xrightarrow{P} 0$$
$$\sup_{\tilde{\theta}\in\Theta_0} |\hat{\mathcal{Q}}_T(\tilde{\theta}) - \mathcal{Q}(\tilde{\theta})| \xrightarrow{P} 0$$

Proof. Parts of this proof are inspired from Freyberger (2015). We prove the 3 statements of the lemma in order

1. Using the properties of the *sup*, the fact that $\forall(A,B)$ rv, $\forall e>0$, $\forall\alpha\in(0,1)$, $\mathbb{P}(A+B>e)\leqslant\mathbb{P}(A>\alpha e)+\mathbb{P}(B>(1-\alpha)e)$ and the previous decomposition of the difference between ξ and $\widehat{\xi}$ we can find an upper bound on the probability that that the difference between $\hat{g}_T(\cdot)$ and $g_T(\cdot)$ is above a deviation: For any $e_1>0$

$$\begin{aligned}
\mathbb{P}(\sup_{\tilde{\theta}} \sqrt{T}||\hat{g}_T(\theta,h) - g_T(\theta,h)||_2 > e_1) &= \mathbb{P}(\sup_{\tilde{\theta}} \sqrt{T}\frac{1}{T}||\sum_{j,t}(\widehat{\xi}_t(f_0(\cdot|\tilde{\lambda}),\tilde{\beta}) - \xi_t(f_0(\cdot|\tilde{\lambda}),\tilde{\beta}))h(z_{jt})||_2 > e_1) \\
&\leqslant \mathbb{P}(\sup_{\tilde{\lambda}} \sqrt{T}||\frac{1}{T}\sum_{j,t}(\rho^{-1}(s_t,x_{2t},f_0(\cdot|\tilde{\lambda}_0)) - D_j(\rho,s_t,\tilde{\lambda}))h(z_{jt})||_2 > \frac{e_1}{3}) \\
&\quad + \mathbb{P}(\sup_{\tilde{\lambda}} \sqrt{T}||\frac{1}{T}\sum_{j,t}(D_j(\rho,s_t,\tilde{\lambda}) - D_j(\rho,\hat{s}_t,\tilde{\lambda}))h(z_{jt})||_2 > \frac{e_1}{3}) \\
&\quad + \mathbb{P}(\sup_{\tilde{\lambda}} \sqrt{T}||\frac{1}{T}\sum_{j,t}(D_j(\rho,\hat{s}_t,\tilde{\lambda}) - D_j(\hat{\rho},\hat{s}_t,\tilde{\lambda}))h(z_{jt})||_2 > \frac{e_1}{3})
\end{aligned}$$

Then we can prove that each element of the upper bound converges to 0

(a) By properties of contractions and using Lemma 2.4 we have

$$|\rho^{-1}(s_t,x_{2t},f_0(\cdot|\tilde{\lambda}_0)) - D_j(\rho,s_t,\tilde{\lambda})| \leqslant \epsilon^H|\rho^{-1}(s_t,x_{2t},f_0(\cdot|\tilde{\lambda}_0)) - \delta_0| \leqslant \epsilon^H\kappa$$

for some constant κ which exists due to the compactness of Λ_0, $\mathcal{X}$ and $\mathcal{B}$. Thus using the iid nature of the data **??**(i), the speed of the NFP algorithm Assumption E(iii), the triangle inequality, Markov inequality and Cauchy-Schwarz inequality

the 1st element converges to 0

$$\begin{aligned}
&\mathbb{P}(\sup_{\tilde{\lambda}} \sqrt{T}||\frac{1}{T}\sum_{j,t}(\rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}_0)) - D_j(\rho, s_t, \tilde{\lambda}))h(z_{jt})||_2 > \frac{e_1}{3}) \\
&\leqslant \mathbb{P}(\sqrt{T}\epsilon^H\kappa||\frac{1}{T}\sum_{j,t}h(z_{jt})||_2 > \frac{e_1}{3}) \leqslant \mathbb{P}(\sqrt{T}\epsilon^H\frac{1}{T}\sum_{j,t}||h(z_{jt})||_2 > \frac{e_1}{3}) \\
&\leqslant \frac{3\kappa}{e_1}\sqrt{T}\epsilon^H\sum_j\sqrt{\mathbb{E}(||h(z_{jt})||_2^2)} \underset{T\to\infty}{\to} 0
\end{aligned}$$

(b) Note that D_j is continuously differentiable in $s \in (0;1)$ so that it is uniformly continuous in s. Indeed C is $\mathcal{C}^\infty$ in s so that

$$\frac{\partial D(\rho, s_t, \tilde{\lambda})}{\partial s} = \prod_{h=1}^{H}\frac{\partial C(C^{(h-1)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\tilde{\lambda})), s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial s}$$

Next because Λ_0 is compact it can be covered by some finite union of closed balls in $\mathbb{R}^{K_2}$, ie $\Lambda_0 \subset \cup_{c=1}^N \Lambda_{0,c}^N$ with $\forall c = 1, \ldots, N \;\; \Lambda_{0,c}^N = \{\tilde{\lambda} : ||\tilde{\lambda} - \lambda_c||_2 \leqslant r_N\}$, $\lambda_c \in \Lambda_0$ and $r_N \underset{N\to\infty}{\to} 0$. Consequently

$$\begin{aligned}
&\mathbb{P}(\sup_{\tilde{\lambda}} \frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho, s_t, \tilde{\lambda}) - D_j(\rho, \hat{s}_t, \tilde{\lambda}))h_E(z_{jt})||_2 > \frac{e_1}{3}) \\
&\leqslant \mathbb{P}(\max_{c=1,\ldots,N}\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho, s_t, \tilde{\theta}) - D_j(\rho, \hat{s}_t, \tilde{\theta}))h_E(z_{jt})||_2 > \frac{e_1}{3}) \\
&\leqslant \sum_{c=1}^N \mathbb{P}(\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho, s_t, \tilde{\lambda}) - D_j(\rho, \hat{s}_t, \tilde{\lambda})| \; ||h_E(z_{jt})||_2 > \frac{e_1}{3}) \\
&\leqslant \sum_{c=1}^N \mathbb{P}(\frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho, s_t, \lambda_c) - D_j(\rho, \hat{s}_t, \lambda_c))h_E(z_{jt})||_2 > \frac{e_1}{9}) \\
&\quad + \sum_{c=1}^N \mathbb{P}(\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho, s_t, \tilde{\lambda}) - D_j(\rho, s_t, \lambda_c)| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9}) \\
&\quad + \sum_{c=1}^N \mathbb{P}(\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho, \hat{s}_t, \lambda_c) - D_j(\rho, \hat{s}_t, \tilde{\lambda})| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9})
\end{aligned}$$

where the last inequality was obtained using the triangle inequality. Then by uniform continuity of D_j in s it follows that $\exists e_2 > 0$ such that $\forall c \;\; \frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho, s_t, \lambda_c) - D_j(\rho, \hat{s}_t, \lambda_c))h_E(z_{jt})||_2 > \frac{e_1}{9}$ implies $\frac{1}{\sqrt{T}}||\sum_{j,t}(s_t - \hat{s}_t)||_2 > e_2$ thence letting $\mathbb{P}^* = \mathbb{P}(\cdot|n_t, x_t, \xi_t)$

$$\mathbb{P}^*(\frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho,s_t,\lambda_c)-D_j(\rho,\hat{s}_t,\lambda_c))h_E(z_{jt})||_2>\frac{e_1}{9})\leqslant\mathbb{P}^*(\frac{1}{\sqrt{T}}||\sum_{j,t}(s_t-\hat{s}_t)||_2>e_2)$$
$$\leqslant\frac{J\sum_t\mathbb{E}^*(||s_t-\hat{s}_t||_2)}{e_2\sqrt{T}}=\frac{J\sum_t\mathbb{E}^*\left(\sqrt{\sum_j(s_{jt}-\hat{s}_{jt})^2}\right)}{e_2\sqrt{T}}\leqslant\frac{J\sum_t\sqrt{\sum_j\mathbb{E}^*\left((s_{jt}-\hat{s}_{jt})^2\right)}}{e_2\sqrt{T}}$$
$$\leqslant\frac{J\sum_t\sqrt{\sum_j\mathbb{E}^*\left((\frac{1}{n_t}\sum_{i=1}^{n_t}y_{ijt}-\mathbb{E}^*(y_{ijt}))^2\right)}}{e_2\sqrt{T}}=\frac{J\sum_t\sqrt{\sum_j Var^*(\frac{1}{n_t}\sum_{i=1}^{n_t}y_{ijt})}}{e_2\sqrt{T}}$$
$$\leqslant\frac{J\sum_t\sqrt{\sum_j\frac{1}{n_t}Var^*(y_{ijt})}}{e_2\sqrt{T}}\leqslant\frac{J^{3/2}}{e_2}\frac{1}{\sqrt{T}}\sum_t\frac{1}{\sqrt{n_t}}$$

where Markov inequality, Jensen inequality, the fact that $y_{ijt}\in\{0;1\}$, that ε_{ijt} is iid extreme-value type 1 distributed across i, j and t, and the fact that n_t is iid and independent of all other variables have been used. Then taking the expectations and summing over N on both sides implies by Assumption E(i)

$$\sum_{c=1}^{N}\mathbb{P}(\frac{1}{\sqrt{T}}||\sum_{j,t}(D_j(\rho,s_t,\lambda_c)-D_j(\rho,\hat{s}_t,\lambda_c))h_E(z_{jt})||_2>\frac{e_1}{9})\leqslant\frac{J^{3/2}N}{e_2}\sqrt{T}\mathbb{E}(n_t^{-1/2})\underset{T\to\infty}{\to}0$$

Next using continuity of D_j in $\tilde{\lambda}$ it must be that for any $e_1>0$ there exists some N such that $\forall\tilde{\lambda}\in\Lambda_{0,c}^N$ such that $||\tilde{\lambda}-\lambda_c||_2\leqslant r_N$ implies

$$\frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho,s_t,\tilde{\lambda})-D_j(\rho,s_t,\lambda_c)|\;||h_E(z_{jt}||_2\leqslant\frac{e_1}{9}$$

because $r_N\underset{N\to\infty}{\to}0$. By definition of the supremum it also implies that

$$\underset{\tilde{\lambda}\in\Lambda_{0,c}}{sup}\frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho,s_t,\tilde{\lambda})-D_j(\rho,s_t,\lambda_c)|\;||h_E(z_{jt}||_2\leqslant\frac{e_1}{9}$$

The contraposition is that

$$\underset{\tilde{\lambda}\in\Lambda_{0,c}}{sup}\frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho,s_t,\tilde{\lambda})-D_j(\rho,s_t,\lambda_c)|\;||h_E(z_{jt}||_2>\frac{e_1}{9}$$

implies $\forall\tilde{\lambda}\in\Lambda_{0,c}^N\quad||\tilde{\lambda}-\lambda_c||_2>r_N$ which is impossible by definition of $\Lambda_{0,c}^N$. Consequently

$$\sum_{c=1}^{N}\mathbb{P}(\underset{\tilde{\lambda}\in\Lambda_{0,c}}{sup}\frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho,s_t,\tilde{\lambda})-D_j(\rho,s_t,\lambda_c)|\;||h_E(z_{jt}||_2>\frac{e_1}{9})$$
$$\leqslant\sum_{c=1}^{N}\mathbb{P}(\cap_{\tilde{\lambda}\in\Lambda_{0,c}^N}||\tilde{\lambda}-\lambda_c||_2>r_N)=0$$

Similarly

$$\sum_{c=1}^{N} \mathbb{P}\Big(\sup_{\tilde{\lambda}\in\Lambda_{0,c}} \frac{1}{\sqrt{T}} \sum_{j,t} |D_j(\rho, \hat{s}_t, \tilde{\lambda}) - D_j(\rho, \hat{s}_t, \lambda_c)| \; ||h_E(z_{jt}||_2 > \frac{e_1}{9}\Big) = 0$$

(c) With the same arguments as in (b)

$$\begin{aligned}
&\mathbb{P}(\sup_{\tilde{\lambda}} \frac{1}{\sqrt{T}} || \sum_{j,t} (D_j(\rho, \hat{s}_t, \tilde{\lambda}) - D_j(\hat{\rho}, \hat{s}_t, \tilde{\lambda})) h_E(z_{jt}) ||_2 > \frac{e_1}{3}) \\
\leqslant & \sum_{c=1}^{N} \mathbb{P}(\frac{1}{\sqrt{T}} || \sum_{j,t} (D_j(\rho, \hat{s}_t, \lambda_c) - D_j(\hat{\rho}, \hat{s}_t, \lambda_c)) h_E(z_{jt}) ||_2 > \frac{e_1}{9}) \\
&+ \sum_{c=1}^{N} \mathbb{P}\Big(\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}} \sum_{j,t} |D_j(\rho, \hat{s}_t, \tilde{\lambda}) - D_j(\rho, \hat{s}_t, \lambda_c)| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9}\Big) \\
&+ \sum_{c=1}^{N} \mathbb{P}\Big(\sup_{\tilde{\lambda}\in\Lambda_{0,c}^N} \frac{1}{\sqrt{T}} \sum_{j,t} |D_j(\hat{\rho}, \hat{s}_t, \lambda_c) - D_j(\hat{\rho}, \hat{s}_t, \tilde{\lambda})| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9}\Big) \\
= & \sum_{c=1}^{N} \mathbb{P}(\frac{1}{\sqrt{T}} || \sum_{j,t} (D_j(\rho, \hat{s}_t, \lambda_c) - D_j(\hat{\rho}, \hat{s}_t, \lambda_c)) h_E(z_{jt}) ||_2 > \frac{e_1}{9})
\end{aligned}$$

where $D_j(\rho, s_t, \lambda_c) = C^{(H)}(\delta_0, s_t, x_{2t}, f_0(\cdot|\lambda_c))$. D_j is $\mathcal{C}^\infty$ in $\rho \in (0;1)$, moreover $\rho_j(\delta_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ and $\hat{\rho}_j(\delta_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))$ are continuously differentiable in Λ_0. Therefore there exists some $e_2 > 0$ such that

$$\frac{1}{\sqrt{T}} \sum_{j,t} |D_j(\rho, \hat{s}_t, \lambda_c) - D_j(\hat{\rho}, \hat{s}_t, \lambda_c)| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9}$$

implies $\sup_{a\in\mathcal{B}} \frac{1}{\sqrt{T}} \sum_{j,t} ||\rho(a, x_{2t}, f_0(\cdot|\lambda_c) - \hat{\rho}(a, x_{2t}, f_0(\cdot|\lambda_c))||_2 > e_2$, and as $\mathcal{B}$ is compact we can cover it by $\tilde{N}$ closed balls $\mathcal{B}_b^{\tilde{N}} = \{a \in \mathcal{B} : ||a - a_b|| \leqslant r_{\tilde{N}}\}$ with $a_b \in \mathcal{B}$ for any $b = 1, \ldots, \tilde{N}$ so that

$$\begin{aligned}
&\sum_{c=1}^{N} \mathbb{P}(\frac{1}{\sqrt{T}} \sum_{j,t} |D_j(\rho, \hat{s}_t, \lambda_c) - D_j(\hat{\rho}, \hat{s}_t, \lambda_c)| \; ||h_E(z_{jt})||_2 > \frac{e_1}{9}) \\
\leqslant & \sum_{c=1}^{N} \mathbb{P}(\sup_{a\in\mathcal{B}} \frac{1}{\sqrt{T}} \sum_{j,t} ||\rho(a, x_{2t}, f_0(\cdot|\lambda_c) - \hat{\rho}(a, x_{2t}, f_0(\cdot|\lambda_c))||_2 > e_2) \\
\leqslant & \sum_{c,b} \mathbb{P}(\sup_{a\in\mathcal{B}_b^{\tilde{N}}} \frac{1}{\sqrt{T}} \sum_{j,t} ||\rho(a, x_{2t}, f_0(\cdot|\lambda_c) - \hat{\rho}(a, x_{2t}, f_0(\cdot|\lambda_c))||_2 > e_2) \\
= & \sum_{c,b} \mathbb{P}(\frac{1}{\sqrt{T}} \sum_{j,t} ||\rho(a_b, x_{2t}, f_0(\cdot|\lambda_c) - \hat{\rho}(a_b, x_{2t}, f_0(\cdot|\lambda_c))||_2 > e_2)
\end{aligned}$$

where the last equality was obtained reusing arguments from (b). As a consequence let $F_{jt}(v) = \frac{exp(a_{bj}+x'_{2jt}v)}{1+\sum_k exp(a_{bk}+x'_{2kt}v)}$ and $\mathbb{P}^*(\cdot) = \mathbb{P}(\cdot|x_t, \xi_t)$ then using Markov inequality and Cauchy-Schwarz inequality

$$\begin{aligned}
&\mathbb{P}^*(\frac{1}{\sqrt{T}}\sum_{j,t}||\rho(a_b, x_{2t}, f_0(\cdot|\tilde{\lambda})) - \hat{\rho}(a_b, x_{2t}, f_0(\cdot|\tilde{\lambda}))||_2 > e_2) \\
&\leqslant \frac{J\sum_t \mathbb{E}^*(||\hat{\rho}(a_b, x_{2t}, f_0(\cdot|\tilde{\lambda})) - \rho(a_b, x_{2t}, f_0(\cdot|\tilde{\lambda}))||_2)}{e_2\sqrt{T}} \\
&\leqslant \frac{J\sum_t \sqrt{\sum_j \mathbb{E}^*\left((\frac{1}{R}\sum_{r=1}^R F_{jt}(v_R) - \mathbb{E}^*(F_{jt}(v_R)))^2\right)}}{e_2\sqrt{T}} = \frac{J\sum_t \sqrt{\sum_j Var^*(\frac{1}{R}\sum_{r=1}^R F_{jt}(v_r))}}{e_2\sqrt{T}} \\
&\leqslant \frac{J^{3/2}}{e_2}\sqrt{\frac{T}{R}}
\end{aligned}$$

where the fact that v_r are iid draws from $f_0(\cdot|\tilde{\lambda})$ independent from all other variables has been used. It follows by taking the expectation and summing over N and $\tilde{N}$ that

$$\mathbb{P}(\sup_{\tilde{\lambda}} \frac{1}{\sqrt{T}}\sum_{j,t}|D_j(\rho, \hat{s}_t, \tilde{\lambda}) - D_j(\hat{\rho}, \hat{s}_t, \tilde{\lambda})|\ h_E(z_{jt})||_2 \underset{T\to\infty}{\to} 0$$

by Assumption E(i).

2. The 2nd statement is not formally proven as it largely builds on the proof of the 1st statement. To see why recall that

$$\hat{\Gamma}_T(\tilde{\theta}, h) - \Gamma_T(\tilde{\theta}, h) = \frac{1}{T}\sum_{jt} h(z_{jt})\frac{\partial}{\partial\theta}(\widehat{\xi}(\tilde{\theta}) - \xi_{jt}(\tilde{\theta}))'$$

More precisely let $e'_j = (0\ \dots\ 0\ \underbrace{1}_{\text{j-th coordinate}}\ 0\ \dots\ 0)$ then

$$\frac{\partial\xi_{jt}(\tilde{\theta})}{\partial\beta} = -x_{1jt}, \quad \frac{\partial}{\partial\lambda}\xi_{jt}(\tilde{\theta}) = -e'_j\left(\frac{\partial\rho(\delta_t(\tilde{\lambda}), x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial\delta}\right)^{-1}\int\frac{exp(\delta_{jt}(\tilde{\lambda}) + x'_{2jt}v)}{1+\sum_{k=1}^J exp(\delta_{kt}(\tilde{\lambda}) + x'_{2kt}v)}\frac{\partial}{\partial\lambda}f_0(v|\tilde{\lambda})dv$$

Thus the columns of the matrix $\hat{\Gamma}_T(\tilde{\theta}, h) - \Gamma_T(\tilde{\theta}, h)$ associated to the derivative in β are equal to 0. Furthermore using an uniform continuity argument $\left|\frac{\partial\hat{\xi}_{jt}(\tilde{\theta})}{\partial\lambda} - \frac{\partial\xi_{jt}(\tilde{\theta})}{\partial\lambda}\right| > e_1$ is implied by $||\widehat{\delta}_t(\tilde{\lambda}) - \delta_t(\tilde{\lambda})||_2 > e_2$ for some $e_2 > 0$. Using the compactness of Λ_0 and Assumption E it is straightforward that $\sup_{\tilde{\lambda}} ||\hat{\Gamma}_T(\tilde{\theta}, h) - \Gamma_T(\tilde{\theta}, h)||_2 \xrightarrow{P} 0$ for any h which satisfies the conditions in Assumption D.

3. The 3rd statement follows from the 1st. Indeed using Cauchy-Schwarz and properties of the supremum

$$\begin{aligned}\sup_{\tilde{\theta}\in\Theta_0} |\hat{\mathcal{Q}}_T(\tilde{\theta}) - \mathcal{Q}_T(\tilde{\theta})| =& |(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E))'\hat{W}(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E)) \\ & - 2(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E))'\hat{W}g_T(\tilde{\theta}, h_E))| \\ \leqslant & \sup_{\tilde{\theta}\in\Theta_0} ||(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E))||_2^2 \bar{\mu}(\hat{W}) \\ & + 2 \sup_{\tilde{\theta}\in\Theta_0} ||(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E))||_2 \sup_{\tilde{\theta}\in\Theta_0} ||g_T(\tilde{\theta}, h_E))||_2 \bar{\mu}(\hat{W})\end{aligned}$$

where $\bar{\mu}(\cdot)$ maps a square matrix towards its maximum eigenvalue. By D(iv) and definition of the L_2 matrix norm, $\bar{\mu}(\hat{W}) \xrightarrow{P} \bar{\mu}(W)$. Then we apply Jennrich's ULLN: the data is iid, Θ_0 is compact, and $g_T(\tilde{\theta}, h_E) = \sum_j \xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) h_E(z_{jt})$ has an enveloppe with finite absolute 1st moment because $\xi_{jt}(f_0(\cdot|\tilde{\lambda}), \tilde{\beta}) = \rho^{-1}(s_t, x_{2t}, \tilde{\lambda}) - x'_{1jt}\tilde{\beta}$ and $\rho^{-1}(\cdot)$ has a maximum because it is continuous and its input are in a compact and because $\tilde{\beta}$ is in a compact and x_{1jt} has finite 4th moments, see Assumption B; Thus by the CMT $\sup_{\tilde{\theta}\in\Theta_0} ||g_T(\tilde{\theta}, h_E))||_2 \xrightarrow{P} \sup_{\tilde{\theta}\in\Theta_0} ||g(\tilde{\theta}, h_E)||_2$; Finally using the 1st statement we have $||(\hat{g}_T(\tilde{\theta}, h_E) - g_T(\tilde{\theta}, h_E))||_2 \xrightarrow{P} 0$ therefore by the CMT

$$\sup_{\tilde{\theta}\in\Theta_0} |\hat{\mathcal{Q}}_T(\tilde{\theta}) - \mathcal{Q}_T(\tilde{\theta})| \xrightarrow{P} 0$$

□

B.4.5 Asymptotic Properties of the BLP estimator

Lemma 2.6 (Consistency of BLP estimator)
Given parametric assumption $\mathcal{F}_0$ and under assumptions B-E

$$\hat{\theta} \xrightarrow{P} \theta_0$$

Proof. We prove consistency using arguments for the consistency of M-estimators. For any $e_1 > 0$ such that $|\hat{\theta} - \theta_0| > e_1$ then by Assumption D(iii) there exists some $e_2 > 0$ such that $\mathcal{Q}(\hat{\theta}) - \mathcal{Q}(\theta_0) > e_2$ as θ_0 is the unique minimizer of the objective. Thence for any $e_1 > 0$, $\exists e_2 >$

0 such that

$$\begin{aligned}\mathbb{P}(|\hat{\theta}-\theta_0|>e_1) &\leqslant \mathbb{P}(\mathcal{Q}(\hat{\theta})-\mathcal{Q}(\theta_0)>e_2)\\ &= \mathbb{P}(\hat{\mathcal{Q}}_T(\theta_0)-\mathcal{Q}(\theta_0)+\mathcal{Q}(\hat{\theta})-\hat{\mathcal{Q}}_T(\hat{\theta})+\hat{\mathcal{Q}}_T(\hat{\theta})-\hat{\mathcal{Q}}_T(\theta_0)>e_2)\\ &\leqslant \mathbb{P}(\hat{\mathcal{Q}}_T(\theta_0)-\mathcal{Q}(\theta_0)+\mathcal{Q}(\hat{\theta})-\hat{\mathcal{Q}}_T(\hat{\theta})>e_2)\\ &\leqslant \mathbb{P}(\hat{\mathcal{Q}}_T(\theta_0)-\mathcal{Q}(\theta_0)>(1-\alpha)e_2)+\mathbb{P}(\mathcal{Q}(\hat{\theta})-\hat{\mathcal{Q}}_T(\hat{\theta})>\alpha e_2)\end{aligned}$$

where $\alpha \in (0;1)$, the 2nd inequality comes from the fact that $\hat{\mathcal{Q}}_T(\hat{\theta})-\hat{\mathcal{Q}}_T(\theta_0)$ is almost surely negative by definition of $\hat{\theta}$, and the 3rd inequality is obtained by utilizing properties of indicator functions. Then by a direct implication of Lemma 2.5 the right-hand-side converges to 0.

□

Lemma 2.7 (Asymptotic normality of BLP estimator)
Given parametric assumption $\mathcal{F}_0$, under assumptions B-E and under $H_0 : f \in \mathcal{F}_0$

$$\sqrt{T}(\hat{\theta}-\theta_0) = \left(\Gamma'(\theta_0,h_E)W\Gamma(\theta_0,h_E)\right)^{-1}\sqrt{T}\Gamma'(\theta_0,h_E)Wg_T(\theta_0,h_E)+o_P(1)$$

Furthermore under $H_0; f \in \mathcal{F}_0$

$$\begin{aligned}\sqrt{T}(\hat{\theta}-\theta_0) \xrightarrow{d}\mathcal{N}(0,&(\Gamma'(\theta_0,h_E)W\Gamma(\theta_0,h_E))^{-1}\Gamma'(\theta_0,h_E)W\Omega(\mathcal{F}_0,h_E)W\Gamma(\theta_0,h_E)\\ &(\Gamma'(\theta_0,h_E)W\Gamma(\theta_0,h_E))^{-1})\end{aligned}$$

Proof. We prove asymptotic normality using arguments from M-estimators asymptotics. From Taylor's Theorem there exists some $\tilde{\theta}$ such that $||\tilde{\theta}-\theta_0||_2 \leqslant ||\hat{\theta}-\theta_0||_2$ and

$$\begin{aligned}&\hat{g}_T(\hat{\theta},h_E) = \hat{g}_T(\theta_0,h_E)+\hat{\Gamma}_T(\tilde{\theta},h_E)(\hat{\theta}-\theta_0)\\ &\Rightarrow \sqrt{T}\hat{\Gamma}'_T(\hat{\theta},h_E)\hat{W}\hat{g}_T(\hat{\theta},h_E) = \sqrt{T}\hat{\Gamma}'_T(\hat{\theta},h_E)\hat{W}\hat{g}_T(\theta_0,h_E)+\hat{\Gamma}'_T(\hat{\theta},h_E)\hat{W}\hat{\Gamma}_T(\tilde{\theta},h_E)\sqrt{T}(\hat{\theta}-\theta_0)=0\\ &\Leftrightarrow \sqrt{T}(\hat{\theta}-\theta_0) = -\left(\hat{\Gamma}'_T(\hat{\theta},h_E)\hat{W}\hat{\Gamma}_T(\tilde{\theta},h_E)\right)^{-1}\sqrt{T}\hat{\Gamma}'_T(\hat{\theta},h_E)\hat{W}\hat{g}_T(\theta_0,h_E)\end{aligned}$$

where the 1st implication is due to the FOC Assumption D(v). Then, we apply the CMT to $(A,B)\mapsto(A'BA)^{-1}A'B$ which is a continuous mapping if A and B are full rank so that when taking $A=\hat{\Gamma}_T(\hat{\theta},h_E)$ and $B=\hat{W}$ we obtain:

$$\sqrt{T}(\hat{\theta}-\theta_0) = -\left(\Gamma'(\theta_0,h_E)W\Gamma(\theta_0,h_E)\right)^{-1}\sqrt{T}\Gamma'(\theta_0,h_E)Wg_T(\theta_0,h_E)+o_P(1)$$

To prove that $plim\ \hat{\Gamma}_T(\hat{\theta},h_E) = plim\ \hat{\Gamma}_T(\tilde{\theta},h_E) = \Gamma(\theta_0,h_E)$ we make the following decomposition

$$\hat{\Gamma}_T(\hat{\theta},h_E)-\Gamma(\theta_0,h_E) = \hat{\Gamma}_T(\hat{\theta},h_E)-\Gamma_T(\hat{\theta},h_E)+\Gamma_T(\hat{\theta},h_E)-\Gamma(\hat{\theta},h_E)+\Gamma(\hat{\theta},h_E)-\Gamma(\theta_0,h_E)$$

where the 1st difference is $o_P(1)$ by Lemma 2.5, the 3rd difference is $o_P(1)$ by the CMT and the consistency of $\hat{\theta}$, see Lemma 2.6, and the 2nd difference is $o_P(1)$ by Jennrich's ULLN. The ULLN can be applied if and only if $\sum_j h_E(z_{jt}) \frac{\partial \xi_{jt}(\theta)}{\partial \theta}$ has an enveloppe with finite 1st absolute moments: $\xi_{jt}(\theta) = \rho^{-1}(s_t, x_{2t}, f_0(\cdot|\lambda)) - x'_{1jt}\beta$ and $\frac{\partial \xi_{jt}(\theta)}{\partial \beta} = x_{1jt}$ with x_{1jt} has finite moments of order 4 by Assumption B(iv), whereas $\frac{\partial \xi_{jt}(\theta)}{\partial \lambda} = \frac{\partial \rho^{-1}(s_t, x_{2t}, f_0(\cdot|\tilde{\lambda}))}{\partial \lambda}$ and ρ^{-1} is C^∞ with arguments (s_t, x_{2t}, λ) which take values in a compact thus $\frac{\partial \rho^{-1}}{\partial \lambda}$ has bounds.

Thence $plim\ \hat{\Gamma}_T(\hat{\theta}, h_E) = plim\ \hat{\Gamma}_T(\tilde{\theta}, h_E) = \Gamma(\theta_0, h_E)$ which is full rank by Assumption D(ii), $plim\ \hat{W} = W$ which is full rank by Assumption D(iv), and by Lemma 2.5 $plim\ \sqrt{T}(\hat{g}_T(\theta_0, h_E) - g_T(\theta_0, h_E)) = 0$ so we can apply the aforementioned CMT and by the CLT which can be applied because $g(\theta_0, h_E) = 0$ under the null

$$\begin{aligned}\sqrt{T}(\hat{\theta} - \theta_0) &= -\left(\Gamma'(\theta_0, h_E) W \Gamma(\theta_0, h_E)\right)^{-1} \sqrt{T} \Gamma'(\theta_0, h_E) W g_T(\theta_0, h_E) + o_P(1) \\ &\xrightarrow{d} \mathcal{N}(0, (\Gamma'(\theta_0, h_E) W \Gamma(\theta_0, h_E))^{-1} \Gamma'(\theta_0, h_E) W \Omega(\mathcal{F}_0, h_E) W \Gamma(\theta_0, h_E) \\ &\quad (\Gamma'(\theta_0, h_E) W \Gamma(\theta_0, h_E))^{-1})\end{aligned}$$

□

B.4.6 Asymptotic distribution of the test statistic

Proof of Theorem 5.1

Proof. This proof leans heavily on the proof of Lemma 2.7. By Taylor's Theorem there exists $\tilde{\theta}$ such that $||\tilde{\theta} - \theta_0||_2 \leqslant ||\hat{\theta} - \theta_0||_2$

$$\begin{aligned}\sqrt{T}\hat{g}_T(\hat{\theta}, h_D) &= \sqrt{T}\hat{g}_T(\theta_0, h_D) + \hat{\Gamma}_T(\tilde{\theta}, h_D)\sqrt{T}(\hat{\theta} - \theta_0) \\ &= (I_{|h_D|_0} \quad - \Gamma(\theta_0, h_D)(\Gamma'(\theta_0, h_D) W \Gamma(\theta_0, h_D))^{-1} \Gamma'(\theta_0, h_D) W) \sqrt{T} \begin{pmatrix} g_T(\theta_0, h_D) \\ g_T(\theta_0, h_E) \end{pmatrix} + o_P(1) \\ &\equiv (I_{|h_D|_0} \quad G) \sqrt{T} \begin{pmatrix} g_T(\theta_0, h_D) \\ g_T(\theta_0, h_E) \end{pmatrix} + o_P(1)\end{aligned}$$

The second equality is obtained by relying on the proof of Lemma 2.7 to express $\sqrt{T}(\hat{\theta} - \theta_0)$ as a function of moments, by relying on Lemma 2.5 so that $plim\ \sqrt{T}\hat{g}_T(\theta_0, h_D) = plim\ \sqrt{T}g_T(\theta_0, h_D)$ and $plim\ \hat{\Gamma}_T(\tilde{\theta}, h_D) = plim\ \Gamma_T(\theta_0, h_D)$, and by using the CMT.

- Under $H_0 : f \in \mathcal{F}_0$ then $\mathbb{E}\left[\sum_j h_D(z_{jt})\xi_{jt}(\theta_0)\right] = 0$ by LIE. So using the CLT and Slutsky's Lemma we obtain

$$\sqrt{T}\hat{g}_T(\hat{\theta}, h_D) \xrightarrow{d} Z \sim \mathcal{N}(0, \Omega_0)$$

where

$$\Omega_0 = \begin{pmatrix} I_{|h_D|_0} & G \end{pmatrix} \begin{pmatrix} \Omega(\mathcal{F}_0, h_D) & \Omega(\mathcal{F}_0, h_D, h_E) \\ \Omega(\mathcal{F}_0, h_D, h_E)' & \Omega(\mathcal{F}_0, h_E) \end{pmatrix} \begin{pmatrix} I_{|h_D|_0} \\ G' \end{pmatrix}$$

with

$$\Omega(\mathcal{F}_0, h_D) = \mathbb{E}\left[\left(\sum_j \xi_{jt}(f(.|\lambda_0), \beta_0)h_D(z_{jt})\right)\left(\sum_j h_D(z_{jt})\xi_{jt}(f_0(.|\lambda_0), \beta_0)\right)'\right]$$

$$\Omega(\mathcal{F}_0, h_D, h_E) = \mathbb{E}\left[\left(\sum_j \xi_{jt}(f(.|\lambda_0), \beta_0)h_D(z_{jt})\right)\left(\sum_j h_E(z_{jt})\xi_{jt}(f_0(.|\lambda_0), \beta_0)\right)'\right]$$

$$G = -\Gamma(\theta_0, h_D)\left[\Gamma(\theta_0, h_E)'W\Gamma(\theta_0, h_E)\right]^{-1}\Gamma(\theta_0, h_E)'W$$

Thence by the continuous mapping theorem:

$$S(h_D, \mathcal{F}_0, \hat{\theta}) = \hat{g}_T(\hat{\theta}, h_D)'\hat{\Sigma}\hat{g}_T(\hat{\theta}, h_D) \xrightarrow{d} Z'\Sigma Z$$

□

- Under H_1' : $\mathbb{E}\left[\sum_j h_D(z_{jt})\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)\right] \neq 0$, we have by Lemma 2.5, by consistency of $\hat{\theta} \xrightarrow{P} \theta_0$ and the CMT:

$$\hat{g}_T(\hat{\theta}, h_D) = g_T(\theta_0, h_D) + o_P(1)$$

Thus by Assumption D(iv) and the CMT

$$\frac{S(h_D, \mathcal{F}_0, \hat{\theta})}{T} \xrightarrow{P} \underbrace{\mathbb{E}\left[\sum_j h_D(z_{jt})\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)\right]' \Sigma \mathbb{E}\left[\sum_j h_D(z_{jt})\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)\right]}_{\kappa(h_D, \mathcal{F}_0, \theta_0)}$$

Under H_1', $\kappa(h_D, \mathcal{F}_0, \theta_0)$ is strictly positive because Σ is positive definite. Thence,

$$\begin{aligned} \forall q \in \mathbb{R} \lim_{T\to\infty} \mathbb{P}(S(h_D, \mathcal{F}_0, \hat{\theta}) > q) &= \lim_{T\to\infty} \mathbb{P}\left(\frac{S(h_D, \mathcal{F}_0, \hat{\theta}) - q}{T} > 0\right) \\ &= \mathbb{P}(\kappa(h_D, \mathcal{F}_0, \theta_0) > 0) \\ &= 1 \end{aligned}$$

where the 2nd equality holds because convergence in probability implies convergence in distribution.

□

B.4.7 Application of Theorem 5.1 to the 2 polar cases

1. **Sargan-Hansen J test**
If $h_D = h_E$, with W and Σ are set to be equal to the GMM 2-step optimal weighting matrix

$$\Sigma = W = \mathbb{E}\left[\left(\sum_j \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_E(z_{jt}) \right) \left(\sum_j \xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) h_E(z_{jt}) \right)' \right]^{-1} = \Omega(\mathcal{F}_0, h_E)^{-1}$$

Then under H_0:

$$S(h_D, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} \chi^2_{|h_E|_0 - |\theta|_0}$$

Proof. By applying Theorem 5.1, we have:

$$S(h_D, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} Z'\Sigma Z \text{ with } Z \sim \mathcal{N}(0, \Omega_0)$$

If $h_D = h_E$ and $W = \Omega(\mathcal{F}_0, h_E)^{-1}$ then Ω_0 simplifies to

$$\begin{aligned}\Omega_0 &= \Omega(\mathcal{F}_0, h_E) - \Gamma(\theta_0, h_E) \left[\Gamma(\theta_0, h_E)' \Omega(\mathcal{F}_0, h_E)^{-1} \Gamma(\theta_0, h_E) \right]^{-1} \Gamma(\theta_0, h_E)' \\ &= \Omega(\mathcal{F}_0, h_E)^{1/2} M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)} \Omega(\mathcal{F}_0, h_E)^{1/2}\end{aligned}$$

with $M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)} \equiv I_{|h_E|_0} - P_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)}$ is the orthogonal projection on the space orthogonal to $\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)$. Let $\tilde{Z} \sim \mathcal{N}(0, I_{|h_E|_0})$, we have by definition:

$$\begin{aligned}Z = \Omega(\mathcal{F}_0, h_E)^{1/2} M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)} \tilde{Z} &\implies \Sigma^{1/2} Z = M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)} \tilde{Z} \\ &\implies Z'\Sigma Z = \tilde{Z}' M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)} \tilde{Z}\end{aligned}$$

The second line comes from symmetry and idempotence of $M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)}$. Orthogonal projections have eigenvalues equal to either 0 or 1 with the number of eigenvalues equal to one corresponding to the rank of the space it projects into, which in our case is $|h_E| - |\theta|_0$. If we denote by V the matrix of eigenvectors of $M_{\Omega(\mathcal{F}_0, h_E)^{-1/2}\Gamma(\theta_0, h_E)}$ then note that $V'\tilde{Z} \sim \mathcal{N}(0, I_{|h_E|_0})$ so that

$$Z'\Sigma Z = \sum_{k=1}^{|h_E|_0 - |\theta|_0} (V'\tilde{Z})_k^2 \sim \chi^2_{|h_E|_0 - |\theta|_0}$$

□

2. **Non-overlapping h_D and h_E**
If Ω_0 is full rank and if the econometrician sets $\Sigma = \Omega_0^{-1}$, then our test statistic has the following asymptotic distribution under H_0:

$$S(h_T, \mathcal{F}_0, \hat{\theta}) \xrightarrow{d} \chi^2_{|h_D|_0}$$

One sufficient condition for Ω_0 being full rank is $(\xi_{jt}(f(\cdot|\lambda_0), \beta_0))_{j=1}^J$ is independent across j and $(h_E(z_{jt}), h_D(z_{jt}))$ not being perfectly colinear.

Proof. The asymptotic result is direct; $(\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0))_{j=1}^J$ being independent across j and $(h_E(z_{jt}), h_D(z_{jt}))$ not being perfectly colinear implies that

$$\Omega(\mathcal{F}_0, h_E, h_D) = \sum_j \mathbb{E}\left[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)^2 h_E(z_{jt}) h_D(z_{jt})'\right]$$

$$\Rightarrow \Omega_0 = \sum_j (I_{|h_D|_0} \quad G) Var \left(\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0) \begin{pmatrix} h_D(z_{jt}) \\ h_E(z_{jt}) \end{pmatrix} \right) \begin{pmatrix} I_{|h_D|_0} \\ G' \end{pmatrix}$$

Thus Ω_0 is positive definite because it is the sum of positive definite matrices. □

B.5 Properties of the MPI in the composite specification test: $f \in \mathcal{F}_0$

Proof of Proposition 3.3.

From Corollary **??**. Under Assumption A,

$$\begin{aligned}
H_a : f \notin \mathcal{F}_0 &\implies \mathbb{E}[\xi_{jt}(f_0(\cdot|\lambda_0), \beta_0)|z_{jt}] \neq 0 \ \ a.s \\
&\implies \mathbb{E}[\xi_{jt}(f_0(.|\lambda_0), \beta_0)|z_{jt}]^2 > 0 \ a.s \\
&\implies \mathbb{E}\left[\mathbb{E}[\xi_{jt}(f_0(.|\lambda_0), \beta_0)|z_{jt}]^2\right] > 0 \\
&\implies \mathbb{E}\left[\mathbb{E}[\xi_{jt}(f_0(.|\lambda_0), \beta_0)\mathbb{E}[\xi_{jt}(f_0(.|\lambda_0)|z_{jt}]|z_{jt}]\right] > 0 \\
&\implies \mathbb{E}\left[\xi_{jt}(f_0(.|\lambda_0), \beta_0)\mathbb{E}[\xi_{jt}(f_0(.|\lambda_0)|z_{jt}]\right] > 0 \\
&\implies \forall \alpha \neq 0 \ \ H_1' : \mathbb{E}\left[\xi_{jt}(f_0(.|\lambda_0), \beta_0) \underbrace{\alpha \mathbb{E}[\Delta_{0,a}^{\xi_{jt}}|z_{jt}]}_{h_D^*(z_{jt})}\right] \neq 0
\end{aligned}$$

From Theorem 5.1, under assumptions B-E,

$$H_a \implies \forall q \in \mathbb{R}^+, \ \ \mathbb{P}(S(h_D^*, \mathcal{F}_0, \hat{\theta}) > q) \to 1$$

□

C Additional results and comments

C.1 Literature on the identification of the distribution of RC

In this section, we briefly summarize some recent findings on the identification of random coefficients in multinomial choice models has been extensively studied in the literature. In their seminal paper, Berry and Haile (2014) shows the identification of the demand functions ρ in a framework that encompasses the BLP model but their result does not entail identification of the random coefficients' distribution per se. To achieve their identification result, they require a completeness condition on the instruments as well as additional conditions (eg: connected substitutes) to ensure invertibility of the demand functions. They also need to impose that at least one of the product characteristic has a coefficient that is not random and that is equal to 1. Notice that in BLP model, the structure implied by the logit shock guarantees invertibility of the demand functions.

Fox et al. (2012) provides conditions under which the distribution of random coefficients is identified in a mixed logit model with micro-level data and no endogeneity. Their identification result requires continuous characteristics in x_{2t} and rules out interaction terms (eg polynomial terms of x_{2jt}). Moreover, their result is restricted to distributions of random coefficients with a compact support - excluding for instance a normally distributed random coefficient.

Fox and Gandhi (2011) investigates the identification of the joint distribution of random coefficients v_i and idiosyncratic shocks ε_{ijt} in aggregate demand models without endogeneity. They also consider a setting where endogeneity is introduced in a very restrictive way. They show identification under a special regressor assumption and finite support of the unobserved heterogeneity. The special regressor assumption assumes that a variable in x_{1t} has full support and has an associated coefficient that is either 1 or -1. This special regressor assumption is very common in the literature on the identification of random coefficients (see Ichimura and Thompson (1998), Berry and Haile (2009), Matzkin (2007) and Lewbel (2000)). Their framework does not nest the standard BLP model as ϵ_{ijt} and v_i are both assumed to have a finite support but it is more general in other dimensions. They do not exploit the logit distributional assumption on ε_{ijt}, they do not impose independence between v_i and ε_{ijt}, their identification argument can be extended to the case where multiple goods are purchased.

In a setting much closer to ours, Dunker et al. (2022) studies the identification of the distribution of random coefficients in endogenous aggregate demand models which includes the BLP model as a special case (in particular, no parametric assumption is made on the idiosyncratic shock ε_{ijt}). They make a clever use of the Radon transform to identify f. The price they have to incur for flexibility is that they need to make stringent assumptions on the product characteristics: variables in x_t are required to be continuous and to satisfy a joint full support assumption. The idea is to exploit the variation in the covariates in order to trace out the distribution of rc f. Unfortunately, these requirements are rarely met in real data sets.

In contrast to the rest of the literature, Wang (2022) adopts all the parametric assumptions in the standard BLP model and looks for the set of minimal assumptions under which the distribution of random coefficients is identified. This approach allows him to obtain sufficient conditions which are much less stringent than the rest of the literature (no special regressor assumption, no full support assumption, no continuity assumption). To be more specific, he shows that if the demand functions are identified on an open set of $\mathbb{R}^J$[38], then the distribution of random coefficients is identified. His proof astutely exploits the real analytic property of the demand functions[39].

C.2 Feasible MPI: conditional expectation

C.3 Choice of the large-T asymptotics

In this paper, we study the asymptotics of our test when the number of markets T grows to infinity. We could also develop an asymptotic theory with J growing to infinity and T staying fixed but there are many arguments against it. First, from an economic stand point, it's hard to conceptually think of a market with a very large number of products and some form of independence across products. Second, from a theoretical point of view there is a tension between the identification of demand which require all market shares to be strictly positive, see Berry and Haile (2014), and the large market asymptotics which require all market shares to tend to 0 as J grows to infinity, see Berry et al. (2004). At the same time it is well established that a many (weak) instruments problem can easily occur in a BLP model with a fixed number of markets and many products especially when using the traditional BLP instruments, see Armstrong (2016).

Consequently, only markets with perfect competition and a careful choice of instruments could somehow fit the assumptions necessary for the BLP model to yield consistent estimators and valid tests with large J. Yet in the majority of empirical IO papers the markets have imperfect competition, sometimes oligopolies, and use the traditional BLP instruments. Thus we establish our theory with a large number of independent markets, which is a natural setting for empirical IO papers and which is not plagued with the aforementioned theoretical problems.

C.4 Construction of the interval instruments in practice

We now provide more details on how to construct the interval instruments in practice. The procedure to construct the interval instruments is as follows:

[38] which can be achieved using Theorem 1 in Berry and Haile (2014)

[39] In particular, the real analytic property yields that the local identification of ρ on $\mathcal{D} \subset \mathbb{R}^J$ implies identification of ρ on $\mathbb{R}^J$. From global identification of ρ, he is then able to show that the random coefficients' distribution is identified under a simple rank condition on x_{2t}

1. Given $(\mathcal{F}_0, \hat{W}, h_E)$, the researcher derives the BLP estimator $\hat{\theta}$
2. Then the researcher chooses L points $(v_l)_{l=1}^L \in \mathbb{R}^L$ in the presumed support of $f_0(\cdot|\hat{\lambda})$.
3. Finally, the researcher can construct a set of L interval instruments based on the approximations of the MPI that we develop in sections 4.2 and 4.1.

   - Global approximation: $\{\pi_{j,l}(z_{jt})\}_{l=1,...,L}$ interval instruments which are such that:

$$\mathbb{E}\left[\Delta_j(s_t, x_{2t}, f_0, f_a)|z_{jt}\right] \approx \log\left(\sum_{l=1}^{L} \omega_l \ \pi_{j,l}(z_{jt})\right) \text{ with } \pi_{j,l}(z_{jt}) = \frac{\frac{\exp(x'_{2jt}v_l)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\delta}^0_{kt}+x'_{2kt}v_l\right\}}}{\int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{2jt}v)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\delta}^0_{jt}+x'_{2kt}v\right\}} f_0(v)dv}$$

     with $\hat{\delta}^0_t$ the linear projection of δ^0_t on z_{jt} (or a carefully chosen subset of z_{jt}).

   - Local approximation: $\{\bar{\pi}_{j,l}(z_{jt})\}_{l=1,...,L}$ interval instruments such that

$$\mathbb{E}[\Delta_j(s_t, x_{2t}, f_0, f_a)|z_{jt}] \approx \sum_{l=1}^{L} \bar{\omega}_l \ \bar{\pi}_{j,l}(z_{jt})$$

$$\text{with } \bar{\pi}_{j,l}(z_{jt}) = \left(\frac{\partial \rho(\hat{\delta}^0_t, x_{2t}, f_0)}{\partial \delta}\right)^{-1} \left[\frac{\exp(\hat{\delta}^0_t + x_{2t}v_l)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\delta}^0_{kt}+x'_{2kt}v_l\right\}} - \rho_j(\hat{\delta}^0_t, x_{2t}, f_0)\right]$$

     with $\hat{\delta}^0_t$ the linear projection of δ^0_t on z_{jt} (or a carefully chosen subset of z_{jt}).

Choice of the L points in the domain of f_a The researcher doesn't know a priori the support of the true density f_a. Thus, he/she must choose points in the domain of definition of f_a. If this choice coincides with points of the support where $|f_0(\cdot|\lambda_0) - f_a|$ is large, then this choice generates more informative instruments. In practice, one can take points in the high density regions of $f_0(\cdot|\lambda_0)$ (eg if $\mathcal{F}_0$ is the Gaussian family, then one can take points around the mean λ_0). The choice of of the number of instruments N obeys a usual bias variance tradeoff. On the one hand, a large L allows to better approximate the MPI and thus increases the detection ability of the instruments. On the other hand, it is well-known that a larger number of instruments can induce finite sample bias and can distort asymptotic distributions of estimators and tests such as the over-identification test[40]; For these reasons we advise not to use too few or too many interval instruments, in our simulations and application we use between 10 and 20 instruments. We leave a formal analysis of the optimal choice of L and of the general approximations properties of the interval instruments for future work.

[40] see Roodman (2009) for a review on the effect of many possibly weak moments on estimation and testing

C.5 Feasible MPIs for estimation

As for the global approximation we derived in section 4.2, it is straightforward to show that for any candidate $f_0(\cdot|\lambda_0)$, we can rewrite this approximation of the non-linear part of the MPI as follows:

$$\mathbb{E}[\Delta_j(s_t, x_{2t}, f_0(\cdot|\lambda_0), f_a)|z_{jt}] \approx \log\left(\sum_{l=1}^{L} \bar{\omega}_l(f_0(\cdot|\lambda_0), f_a)\ \hat{\bar{\pi}}_{j,l}(z_{jt})\right) \text{ with } \hat{\bar{\pi}}_{j,l}(z_{jt}) = \frac{\exp(x'_{2jt}v_l)}{1+\sum_{k=1}^{J}\exp\left\{\hat{\bar{\delta}}^a_{jt}+x_{2jk}v_l\right\}}$$

$$\text{and } \bar{\omega}_l(f_0, f_a) = \frac{\bar{\omega}_l(f_a)}{\int_{\mathbb{R}^{K_2}} \frac{\exp(x'_{2jt}v)}{1+\sum_{k=1}^{J}\exp\left\{\delta^0_{jt}+x'_{2jk}v\right\}} f_0(\cdot|\lambda_0)(v)dv}$$

with $\hat{\bar{\delta}}^a_{jt}$ projected first stage estimates of δ^a_{jt}, which can be obtained, for example, under the logit specification. $\hat{\bar{\pi}}_{j,l}(z_{jt})$ don't depend on f_0 and can be used for estimation.

C.6 Estimation procedure when the distribution of RC is a mixture

In this section, we present a procedure to estimate the BLP model when the distribution of RC is parametrized as a mixture. Namely, we perform the estimation under $H_0 : f \in \mathcal{F}_0$ with $\mathcal{F}_0$ the family of Gaussian mixtures with L components. The pdf of a Gaussian mixture writes as follows:

$$\forall x \in \mathbb{R}\ ,\ f_0(x|\lambda_0) = \sum_{l=1}^{L} p_{l0} f_l(x|\lambda_{l0}) \qquad \sum_{l=1}^{L} p_{l0} = 1 \qquad L \geqslant 1$$

where $f_{l0}(\cdot|\lambda_{l0})$ is the pdf of a $\mathcal{N}(\mu_{l0}, \sigma^2_{l0})$.

As long as the means are different ($\mu_{l0} \neq \mu_{l'0}\ \forall l \neq l'$), the gaussian mixture is uniquely characterized by the vector $\lambda_0 = (p_{10}, \ldots, p_{L0}, \mu_{10}, \ldots, \mu_{L0}, \sigma^2_{10}, \ldots, \sigma^2_{L0})$ up to permutations of indexes[41]. The objective of our procedure is to estimate the parameters of the model $\theta_0 = (\beta_0, \lambda_0)$ where λ_0 characterizes the mixture. In general, the problem of estimating a density by a mixture is solved through the use of the well-known Expectation-Maximization (EM) algorithm. In our case, the application of this algorithm is made difficult by two main obstacles. First, we do not observe directly the random coefficients. Second, we do not have individual choice data which would have enabled us to construct a likelihood as in Train (2008). As an alternative, we propose to adapt the BLP estimation procedure to estimate the parameters of a mixture of gaussians instead of the single normal distribution. The mixture affects the derivation of the market shares. The random coefficient v_i is now a gaussian

[41] If for some $l \neq l'$ we have $\mu_{l0} = \mu_{l'0}$ then the Gaussian mixture becomes observationally equivalent to an infinite number of other Gaussian mixtures

mixture. Hence, $v_i = \sum_{l=1}^{L} \mathbf{1}\{D_i = l\} v_{il}$ where $(v_{il})_{i=1}^{n}$ are iid and have density $f_{l0}(\cdot|\lambda_{l0})$ known up to λ_{l0} for $l = 1, \ldots, L$, and where $(D_i)_{i=1}^{n}$ are iid categorically distributed with pmf $\mathbb{P}(D_i = l) = p_{l0}$. For all market t and product j, the demand functions are as follows:

$$\begin{aligned}
\rho_j(\delta_t, x_{2t}, f_0(.|\lambda_0)) &= \mathbb{P}(j \text{ chosen in market } t \text{ by } i | x_{1t}, x_{2t}, \xi_t) \\
&= \int_{\mathbb{R}} \frac{\exp\{x'_{1jt}\beta_0 + x'_{2jt}v + \xi_{jt}\}}{1 + \sum_{j'=1}^{J} \exp\{x'_{1j't}\beta_0 + x'_{2j't}v + \xi_{j't}\}} f_0(v|\lambda_0) dv \\
&= \sum_{l=1}^{L} p_{l0} \int_{\mathbb{R}} \frac{\exp\{\delta_{jt} + x'_{2jt}v\}}{1 + \sum_{j'=1}^{J} \exp\{\delta_{j't} + x'_{2j't}v\}} f_{l0}(v|\lambda_{l0}) dv
\end{aligned}$$

Reparametrization The parameter λ associated with the mixture consists of the means, the standard deviation and the probability of each component. As highlighted by Ketz (2019) in the simple Gaussian case, the way we parametrize the model can greatly affect the asymptotic properties of the estimator as well as the quality of the estimation. In particular, he shows that the standard deviations σ should be reparametrized in order to avoid boundaries issues when σ close to 0. We follow this parametrization and perform the minimization with respect to $\{(+/-)\sqrt{\sigma_l}\}_{l=1}^{L}$ instead and $(\sigma_l)_{l=1}^{L}$ directly. An additional difficulty in the case of mixtures concerns the estimation of the probabilities associated to each component. These probabilities must all be between 0 and 1 and their sum must be equal to 1. To smoothly integrate these constraints, we perform the optimization with respect to $\gamma = (\gamma_2, \ldots, \gamma_L)$ with $p = (p_1, p_2, \ldots, p_L) = (\frac{1}{1+\sum_{l=2}^{L} exp(\gamma_l)}, \frac{exp(\gamma_2)}{1+\sum_{l=2}^{L} exp(\gamma_l)}, \ldots, \frac{exp(\gamma_L)}{1+\sum_{l=2}^{L} exp(\gamma_l)})$.

Estimation details Apart from the modification in the computation of the market shares and the new parametrization of the model, the estimation procedure with a mixture follows closely the traditional one and the parameters of interest are estimated by minimizing a GMM criterion. Let $\mathcal{Q}(\theta)$ the GMM objective function:

$$\mathcal{Q}(\theta) = \hat{\xi}(\theta)' h_E(Z) W h_E(Z)' \hat{\xi}(\theta)$$

We now describe the derivation of the Gradient that we provide to the minimization program.

$$\frac{\partial \mathcal{Q}}{\partial \theta} = 2 \left[\frac{\partial \hat{\xi}(\theta)}{\partial \theta} \right]' h_E(Z) W h_E(Z)' \hat{\xi}(\theta)$$

Where $\frac{\partial \hat{\xi}(\theta)}{\partial \beta} = -x_1$ and where by the implicit function theorem we have $\hat{\rho}_j(\delta_t, x_{2t}, \lambda) - s_{jt} = 0 \ \ \forall j, t$ which implies:

$$\frac{\partial \hat{\xi}(\theta)}{\partial \lambda} = \frac{\partial \hat{\delta}(\theta)}{\partial \lambda} = -\left[\frac{\partial \hat{\rho}(\delta, x_2, \lambda)}{\partial \delta}\right]^{-1} \frac{\partial \hat{\rho}(\delta, x_2, \lambda)}{\partial \lambda}$$

- $\frac{\partial \rho}{\partial \delta}$ is a $JT \times JT$ diagonal by block matrix such that:

$$\frac{\partial \rho_j(\delta_t, x_{2t}, \lambda)}{\partial \delta_{kt}} = \begin{cases} \sum_l p_l \int \mathcal{T}_{jlt}(v)\left(1 - \mathcal{T}_{klt}(v)\right)\phi_l(v)dv & if \quad j = k \\ -\sum_l p_l \int \mathcal{T}_{jlt}(v)\mathcal{T}_{klt}(v)\phi_l(v)dv & if \quad j \neq k \end{cases}$$

  with $\mathcal{T}_{jlt}(v) \equiv \frac{exp\{\delta_{jt} + x'_{2jt}v_l\}}{1+\sum_{j'=1}^{J} exp\{\delta_{j't} + x'_{2j't}v_l\}}$

- $\frac{\partial \rho}{\partial \lambda}$ is a $JT \times (3L-1)$ matrix such that:

$$\frac{\partial \rho_j(\delta_t, x_{2t}, \lambda)}{\partial \mu_l} = p_l \int \mathcal{T}_{jlt}\left(x_{2jt} - \sum_{j'} \mathcal{T}_{j'lt}x_{2j't}\right)\phi(v)dv$$

$$\frac{\partial \rho_j(\delta_t, x_{2t}, \lambda)}{\partial \sigma_l} = p_l \int \mathcal{T}_{jlt}\left(x_{2jt} - \sum_{j'} \mathcal{T}_{j'lt}x_{2j't}\right)v\phi(v)dv$$

$$\frac{\partial \rho_j(\delta_t, x_{2t}, \lambda)}{\partial \gamma_l} = \sum_{l'=1}^{L} \zeta(l, l') \int \mathcal{T}_{jlt}$$

  With $\zeta(l, l') = \frac{-exp(\gamma_l)}{1+\sum_{k\neq 1} exp(\gamma_k)} \times \frac{exp(\gamma_{l'})}{1+\sum_{k\neq 1} exp(\gamma_k)} + \mathbf{1}\{l = l'\}\frac{exp(\gamma_l)}{1+\sum_{k\neq 1} exp(\gamma_k)} = -p_l \times p_{l'} + \mathbf{1}\{l = l'\}p_l$

C.7 Properties of the feasible approximations of the MPI

So far we have studied the properties of the MPI, which is an ideal instrument that cannot be derived in practice. Nevertheless, in light of the previous results, the MPI provides a useful upper bound on the power that can be reached using our specification test. More precisely, the asymptotic slope reached by the MPI can be interpreted as a power envelope on our specification test. Ideally, we want our specification test, with the approximated MPIs as instruments, to achieve slopes close to the ones reached by the MPI. For the sake of exposition, let us assume homoskedasticity. We now distinguish 2 situations.

First, we consider the case where the econometrician tests H_0 against the true alternative $\bar{H}_a : (f, \beta) = (f_a, \beta_a)$. This situation is not interesting in practice as the econometrician usually

doesn't know the true alternative and doesn't want to specify an alternative. Nevertheless, it illustrates that in this specific case, we can (in theory) derive a consistent estimator of the MPI. Indeed, in this particular case, we can directly derive an analytical expression for the correction term $\Delta_{0,a}^{\zeta_{jt}}$ either using its definition or the expression in 4.2. Next, we must to compute the conditional expectation of our the correction term with respect to z_{jt}. This step is quite challenging because the dimension of z_{jt} is large and because the correction term is heavily non-linear and non-separable with respect to the endogenous variables. In theory, a solution is to perform a Sieve non-parametric estimation of the conditional mean and under standard regularity conditions recover a consistent estimator of $\mathbb{E}[\Delta_{0,a}^{\zeta_{jt}}|z_{jt}]$. Unfortunately, the rate of converge will be extremely slow given the dimension of z_{jt} and we don't recommend to do this in practice. Instead, we suggest to use the global approximation and to "exogenize" the endogenous variables by projecting them on the space spanned by a relevant subset of z_{jt}. As we show in the Appendix, this strategy yields an estimator which converges faster to a first order approximation of the MPI.

Second, we consider the more realistic situation where the econometrician tests H_0 against an unspecified alternative. This is the situation of interest in this paper. In this case, we use the interval instruments that we developed in section 4 as an approximation of the MPI. Due to the different layers of approximations which intervene in the construction of these instruments and the absence of knowledge of f_a, it is quite difficult to establish conditions under which these instruments can reach the optimal slope of the MPI. A thorough analysis of the properties of these instruments is beyond the scope of this paper and may constitute an interesting starting point for future research. In the Appendix, we present a preliminary investigation on the theoretical properties of the local approximation. In spite of the lack of theoretical analysis, our Monte Carlo exercises show that the interval instruments perform really well in finite sample.

Approximation properties of the interval instruments The interval instruments, ie the approximation of the MPI denoted as $\hat{h}_T^*$, work well in practice in the sense that they yield a valid test which is powerful. However it is difficult to prove that the speed of divergence of our test when using them is as large as the speed of divergence when using the true MPI without further assumptions. As described in the previous subsection there are 3 levels of approximation to the MPI: First, only the 1st order of the expansion of the difference between the true error and generated error is considered, another term $\mathcal{R}_0$ remains[42]; Second, the conditional expectation with respect to the full set of instruments is approximated via projections; Third, the integral which appears within this 1st order approximation is estimated via a Riemann sum of N points. Consequently, if $\mathcal{R}_0$ is negligible, if N is very large, and if projecting the difference between the generated error and the true error is equivalent to taking

[42] we have obtained the formula of the approximation of the difference between the true error and the generated error up to the second order

its conditional expectation with respect to the instruments, then the slope $C_{\hat{h}_T^*}$ is equal to $C_{h_T^*}$. This result is summarized in the following proposition:

Proposition 3.1
Under Assumption B and C, and assuming strict homoskedasticity $\mathbb{E}(\xi_{t0}\xi_{t0}'|z_t) = I_J$ *then under* $H_1 : f \notin \mathcal{F}_0$ *there exists some sequence* $(\alpha)_{l=1}^N$ *such that*

$$\hat{h}_T^*(z_{jt})'\alpha + err_{jt} + e\tilde{r}r_{jt} \underset{N\to+\infty}{\longrightarrow} h_T^*(z_{jt})$$

almost surely, for some errors $err_{jt} \in \mathbb{R}^N$ *and* $e\tilde{r}r_{jt} \in \mathbb{R}$ *such that* $e\tilde{r}r_{jt} \underset{N\to+\infty}{\overset{as}{\longrightarrow}} 0$. *As a consequence*

$$C_{\hat{h}_T^*} = \mathbb{E}\left(\sum_j \alpha'\hat{h}_T^*(z_{jt})\hat{h}_T^*(z_{jt})'\alpha\right) + error + e\tilde{r}or$$

for some $error \in \mathbb{R}$ *and some* $e\tilde{r}or \in \mathbb{R}$ *such that* $e\tilde{r}or \underset{N\to+\infty}{\overset{as}{\longrightarrow}} 0$. *If* $error = 0$ *then*

$$C_{\hat{h}_T^*} \underset{N\to+\infty}{\longrightarrow} C_{h_T^*} = \mathbb{E}\left(\sum_j \mathbb{E}(\tilde{\Delta}_{jt}(s_t, x_{2t}, \mathcal{F}_0, f)|z_{jt})^2\right)$$

To further comment on this result err_{jt} (*error*) corresponds to the first and second errors of approximations of the MPI described above and $e\tilde{r}r_{jt}$ (*er*$\tilde{r}$*or*) corresponds to the third; On the other hand $(\alpha)_{l=1}^N$ is a sequence of integration weights whose empirical mean converge to 0. In addition there are 2 conditions necessary for $error = 0$. The first and most important one is that $\mathcal{R}_0$ shoud be close to 0, in other words the 1st order approximation should explain most of the difference between the generated error and the true error. The second condition for *error* to be close to 0 is very likely to be satisfied in practice: We need to be able to approximate well the conditional expectation with respect to z_{jt} of the 1st order approximation of $\tilde{\Delta}$. As noted by Reynaert and Verboven (2014), because most product characteristics are uncorrelated with the unobserved product characteristics, using a Sieve estimator of the conditional expectation or a more practical method as is described in our paper or theirs does not seem to make a lot of difference. If these two conditions are satisfied then err_{jt} is small and therefore *error* is small.

C.7.1 Proof of Proposition 3.1

Using the strict homoskedasticity assumption then from **??** we know that

$$C_{h_T^*} = \mathbb{E}(\mathbb{E}(\tilde{\Delta}_t(\mathcal{F}_0, f)|z_t)'\mathbb{E}(\tilde{\Delta}_t(\mathcal{F}_0, f)|z_t)) = \mathbb{E}(h_T^*(z_t)'h_T^*(z_t))$$

Our goal is therefore to prove that under some conditions

$$\lim_{N\to+\infty} C_{\hat{h}_T^*} = C_{h_T^*}$$

and we do so in four steps: First, we prove that there exists some (err_1, err_2, err_3) such that

$$\hat{h}_T^*(z_t)'\alpha + err_{1t} + err_{2t} + err_{3t} \underset{N\to+\infty}{\to} h_T^*(z_t)$$

Second, we show that $err_{3t} \underset{N\to+\infty}{\to} 0$ almost surely; Third, we show that there exists some $\tilde{h}_T^*(z_t)$ and some $(error, er\tilde{r}or)$ such that

$$C_{\hat{h}_T^*} = C_{\tilde{h}_T^*}, \qquad C_{h_T^*} = \alpha'\mathbb{E}(\tilde{h}_T^*(z_t)'\tilde{h}_T^*(z_t))\alpha + error + er\tilde{r}or$$

and $er\tilde{r}or \underset{N\to+\infty}{\overset{as}{\to}} 0$; Fourth we conclude. We prove each point in order:

- Denote and recall

$$\begin{aligned}
\eta_{jt} &= \int \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v)}{1+\sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v)}(f(v) - f_0(v|\lambda_0))dv \\
\hat{\eta}_{jt,l} &= \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v_l)}{1+\sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v_l)} \\
\Rightarrow \hat{\eta}_{jt}'\alpha &= \sum_l \alpha_l \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v_l)}{1+\sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v_l)} \\
M_t(\cdot) &= x_{1t}\left(\mathbb{E}\left[\sum_j x_{1jt}h_E(z_{jt})'\right] W \mathbb{E}\left[\sum_j h_E(z_{jt})x_{1jt}'\right]\right)^{-1} \mathbb{E}\left[\sum_j x_{1jt}h_E(z_{jt})'\right] W \mathbb{E}\left[\sum_j h_E(z_{jt})\cdot\right] \\
\hat{M} &= \hat{x}_1\left[x_1'h_E(z)(h_E(z)'h_E(z))^{-1}h_E(z)'x_1\right]^{-1} x_1'h_E(z)\hat{W}h_E(z)' \\
M_{t,\partial\rho}^{-1} &= \left(\frac{\partial\rho(\delta_t^0, x_{2t}, f_0(\cdot|\lambda_0)}{\partial\delta}\right)^{-1} \\
\hat{M}_{t,\partial\rho}^{-1} &= \left(\frac{\partial\rho(\hat{\delta}_t^0, \hat{x}_{2t}, f_0(\cdot|\hat{\lambda})}{\partial\delta}\right)^{-1}
\end{aligned}$$

  where $(\hat{x}_1, \hat{x}_2, \hat{\delta}^0)$ are transformations of (x_1, x_2, δ^0) (for instance their projection on the instruments) as described in **??**. Then define $\hat{M}_{\partial\rho}$ of dimension $(J\times T)\times(J\times T)$ which is block diagonal with T blocks of dimension $J\times J$ equal to $\hat{M}_{t,\partial\rho}^{-1}$, and define $\hat{\eta}$ which is

the stacked versions of $\hat{\eta}_{jt}$. Consequently

$$\begin{aligned} h_T^*(z_t) = \mathbb{E}(\tilde{\Delta}(\mathcal{F}_0, f)|z_t) &= \mathbb{E}\left((id - M_t)\Delta(s_t, x_{2t}, \mathcal{F}_0, f)|z_t\right) \\ &= \mathbb{E}\left((id - M_t)(M_{t,\partial\rho}^{-1}\eta_t + \mathcal{R}_0)|z_t\right) \\ \hat{h}_T^*(z_t)\alpha &= A_t\hat{h}_T(z)\alpha \\ &= A_t(I_{J\times T} - \hat{M})\hat{\Delta}_N'\alpha \\ &= A_t(I_{J\times T} - \hat{M})\hat{M}_{\partial\rho}^{-1}\hat{\eta}\alpha \end{aligned}$$

where A_t is the matrix which picks the J observations in t, ie A_t is a $J \times (J \times T)$ matrix of zeros except the block from column $(J-1)t+1$ to Jt which is equal to I_J. In other words

$$\begin{aligned} h_T^*(z_t) = \hat{h}_T^*(z_t)\alpha &+ \mathbb{E}\left((id - M_t)\mathcal{R}_0|z_t\right) \\ &+ \left[\mathbb{E}((id - M_t)M_{t,\partial\rho}^{-1}\eta_t|z_t) - A_t(I_{J\times T} - \hat{M})\hat{M}_{\partial\rho}^{-1} \lim_{N\to+\infty} \hat{\eta}\alpha\right] \\ &+ \left[\lim_{N\to+\infty}(\hat{h}_T^*(z_t)\alpha) - \hat{h}_T^*(z_t)\alpha\right] \\ \equiv \hat{h}_T^*(z_t)\alpha &+ err_{1t} + err_{2t} + err_{3t} \end{aligned}$$

- Next clearly if $(v_l, c_{l,N})_{l=1}^N$ are chosen so that $\forall l \;\; v_l$ is in the support of $f(\cdot) - f_0(\cdot|\hat{\lambda})$ and $v_{l+1} - v_{l+1} = c_{l,N} \underset{N\to+\infty}{\to} 0$ then the Riemann sum

$$\hat{\eta}_{jt}'\alpha = \sum_l \alpha_l \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v_l)}{1 + \sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v_l)} = \sum_l \frac{c_{l,N}}{N}(f(v_l) - f_0(v_l|\hat{\lambda})) \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v_l)}{1 + \sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v_l)}$$

converges to $\int \frac{exp(\hat{\delta}_{jt}^0 + x_{2jt}'v)}{1+\sum_{k=1}^J exp(\hat{\delta}_0 + x_{2kt}'v)}(f(v) - f_0(v|\hat{\lambda}))dv$ almost surely when $N \to +\infty$, see arguments for the convergence of Riemann sums. This integral exists by Assumption C(ii) and implicitly by Assumption **??**(i)-(ii). Therefore for any $t \;\; err_{3t} \underset{N\to+\infty}{\to} 0$ almost surely, which corresponds to $e\tilde{r}r_t$ in the proposition.

- Next for a fixed N, by Assumption **??** using the LLN and the CMT, there exists some $\tilde{h}_T^*(z_t)$ which is the "probability limit" of $\hat{h}_T^*(z_t)$ in the sense that

$$\frac{1}{T}S(\hat{h}_T^*, \mathcal{F}_0, \theta_0) = \frac{1}{T}S(\tilde{h}_T^*, \mathcal{F}_0, \theta_0) + o_P(1), \qquad \tilde{h}_T^*(z_t) = (id - \tilde{M})\tilde{M}_{t,\partial\rho}^{-1}\tilde{\eta}_t, \qquad C_{\hat{h}_T^*} = C_{\tilde{h}_T^*}$$

where $BLP(\cdot|z_t)$ is the best linear projection operator and

$$\tilde{M}_t(\cdot) = \tilde{x}_{1t}\left(\mathbb{E}\left[\sum_j x_{1jt}h_E(z_{jt})'\right]W\mathbb{E}\left[\sum_j h_E(z_{jt})x'_{1jt}\right]\right)^{-1}\mathbb{E}\left[\sum_j x_{1jt}h_E(z_{jt})'\right]W\mathbb{E}\left[\sum_j h_E(z_{jt})\cdot\right]$$

$$\tilde{M}_{t,\partial\rho} = \left(\frac{\partial\rho(\tilde{\delta}^0_t, \tilde{x}_{2t}, \mathcal{F}_0, \lambda_0)}{\partial\delta}\right)^{-1}$$

$$\tilde{\eta}_{jt,l} = \frac{exp(\tilde{\delta}^0_{jt} + \tilde{x}'_{2jt}v_l)}{1+\sum_k exp(\tilde{\delta}^0_{kt} + \tilde{x}'_{2kt}v_l)}$$

$$\tilde{\delta}^0_t = \delta^0_t - BLP(\delta^0_t|z_t)$$

$$\tilde{x}_{1t} = x_{1t} - BLP(x_{1t}|z_t)$$

$$\tilde{x}_{2t} = x_{2t} - BLP(x_{2t}|z_t)$$

As a consequence $h^*_T(z_t)$ rewrites

$$\begin{aligned}
h^*_T(z_t) =& \tilde{h}^*_T(z_t)\alpha + \mathbb{E}\left((id - M_t)\mathcal{R}_0|z_t\right) \\
&+ \left[\mathbb{E}((id-M_t)M^{-1}_{t,\partial\rho}\eta_t|z_t) - (id-\tilde{M}_t)(\tilde{M}^{-1}_{t,\partial\rho}\lim_{N\to+\infty}\tilde{\eta}_t\alpha)\right] \\
&+ \left[\lim_{N\to+\infty}(\tilde{h}^*_T(z_t)\alpha) - \tilde{h}^*_T(z_t)\alpha\right] \\
\equiv& \tilde{h}^*_T(z_t)\alpha + e\tilde{r}r_{1t} + e\tilde{r}r_{2t} + e\tilde{r}r_{3t} \\
\Rightarrow C_{h^*_T(z_t)} \equiv& \alpha'\mathbb{E}(\tilde{h}^*_T(z_t)'\tilde{h}^*_T(z_t))\alpha + er\tilde{r}or + error
\end{aligned}$$

where $er\tilde{r}or$ is a function of $e\tilde{r}r_{3t}$ and therefore converges to 0 almost surely as $N \to +\infty$ and $error$ is a function of $\tilde{h}^*_T(z_t)$, $e\tilde{r}r_{1t}$ and $e\tilde{r}r_{2t}$.

- From the previous point if $e\tilde{r}r_{1t} = e\tilde{r}r_{2t} = 0$, ie

$$\mathcal{R}_0 = 0, \qquad \left[\mathbb{E}((id-M_t)M^{-1}_{t,\partial\rho}\eta_t|z_t) - (id-\tilde{M}_t)(\tilde{M}^{-1}_{t,\partial\rho}\lim_{N\to+\infty}\tilde{\eta}_t\alpha)\right] = 0$$

Then $h^*_T(z_t) \underset{N\to+\infty}{=} \tilde{h}^*_T(z_t)\alpha$ thus $C_{h^*_T} \underset{N\to+\infty}{=} C_{\tilde{h}^*_T\alpha} = C_{\hat{h}^*_T\alpha}$. Finally using the properties of best linear projections it can be shown that $C_{\hat{h}^*_T} = C_{\tilde{h}^*_T} \geqslant C_{\tilde{h}^*_T\alpha} = C_{\hat{h}^*_T\alpha}$ so that $\lim_{N\to+\infty} C_{\hat{h}^*_T} = C_{h^*_T\alpha}$ because $C_{h^*_T\alpha}$ also constitutes an upper bound on $C_{\hat{h}^*_T}$. Indeed

$$\begin{aligned}
C_{\tilde{h}^*_T} &= \mathbb{E}\left(\tilde{\Delta}_t(\mathcal{F}_0,f)'\tilde{h}^*_T(z_t)\right)\mathbb{E}\left(\tilde{h}^*_T(z_t)'\xi_{0t}\xi'_{0t}\tilde{h}^*_T(z_t)\right)^{-1}\mathbb{E}\left(\tilde{h}^*_T(z_t)'\tilde{\Delta}_t(\mathcal{F}_0,f)\right) \\
&= \mathbb{E}\left(\tilde{\Delta}_t(\mathcal{F}_0,f)'\tilde{h}^*_T(z_t)\right)\mathbb{E}\left(\tilde{h}^*_T(z_t)'\tilde{h}^*_T(z_t)\right)^{-1}\mathbb{E}\left(\tilde{h}^*_T(z_t)'\tilde{\Delta}_t(\mathcal{F}_0,f)\right) \\
&\geqslant C_{\tilde{h}^*_T\alpha} = \mathbb{E}\left(\tilde{\Delta}_t(\mathcal{F}_0,f)'\tilde{h}^*_T(z_t)\right)\alpha\mathbb{E}\left(\alpha'\tilde{h}^*_T(z_t)'\tilde{h}^*_T(z_t)\alpha\right)^{-1}\alpha\mathbb{E}\left(\tilde{h}^*_T(z_t)'\tilde{\Delta}_t(\mathcal{F}_0,f)\right)
\end{aligned}$$

where the first second equality is due to the fact that we assume strict exogeneity $\mathbb{E}(\xi_{0t}\xi_{0t}|z_t) = I_J$, and the inequality is due to the fact that the best linear projection of $\tilde{\Delta}_t(\mathcal{F}_0, f)$ on the subspace $\tilde{h}_T^*(z_t)\alpha$ always has lower second moment compared to the best linear projection of $\tilde{\Delta}_t(\mathcal{F}_0, f)$ on the space $\tilde{h}_T^*(z_t)$.

D Monte Carlo experiments

D.1 Counterfactuals under an alternative distribution

Expressions for price and cross-price elasticities as a function of p_1 in the simulation exercise presented in section 6.2

- Price elasticity:

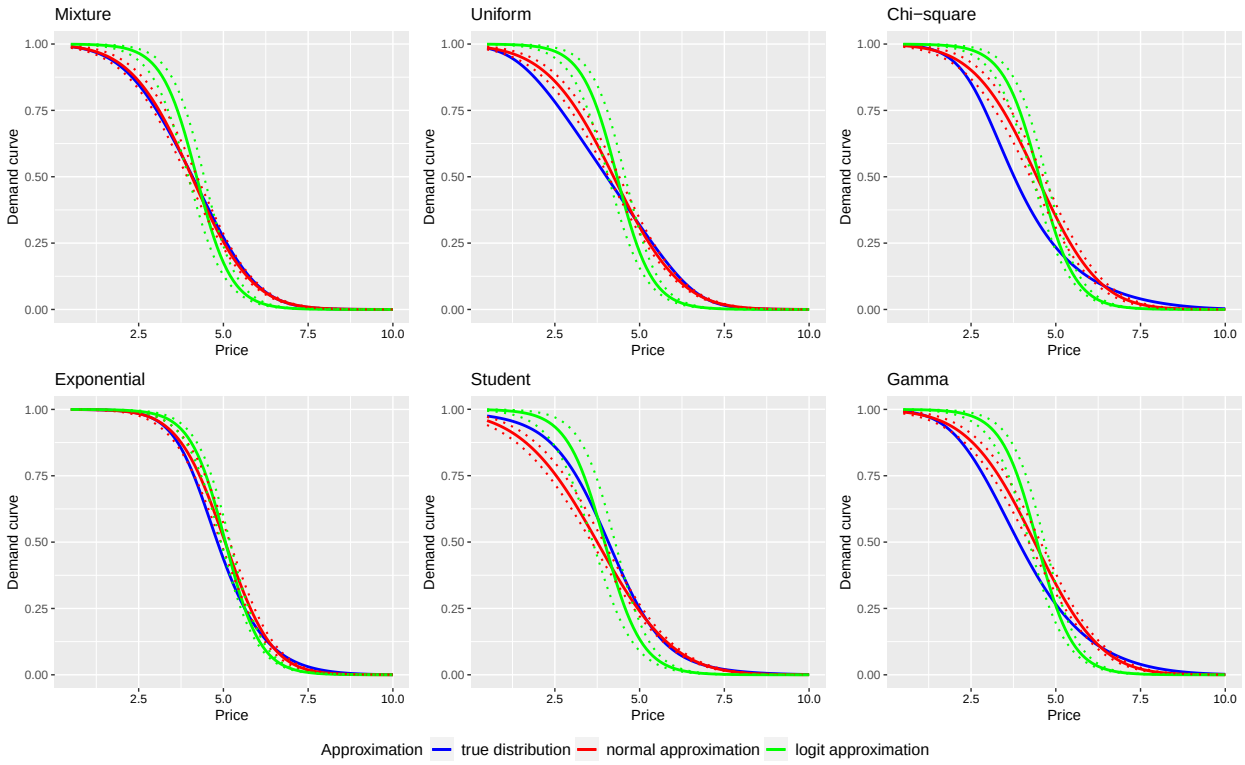
$$\xi_1 = \frac{p_1}{s_1}\frac{\partial s_1}{\partial p_1} = \int -\alpha \left(1 - \frac{\exp\{u_{i1}\}}{1 + \sum_{j' \in \{1,2\}} \exp\{u_{ij'}\}}\right) \frac{exp\{u_{i1}\}}{1 + \sum_{j' \in \{1,2\}} \exp\{u_{ij'}\}} f_\theta(v) dv \phi(\alpha) d\alpha$$

- Cross price elasticity:

$$\xi_{2/1} = \frac{p_1}{s_2}\frac{\partial s_2}{\partial p_1} = \int \alpha \left(\frac{\exp\{u_{i1}\}}{1 + \sum_{j' \in \{1,2\}}^{J} \exp\{u_{ij'}\}}\right) \frac{\exp\{u_{i2}\}}{1 + \sum_{j' \in \{1,2\}}^{J} \exp\{u_{ij'}\}} f_\theta(v) dv \phi(\alpha) d\alpha$$

Here, we plot the demand functions generates by the different approximations of the true densities

Figure 9: Demand function



D.2 Finite sample performance of the test

Practical implementation of the test For each setting, we estimate the model for 1000 replications. Minimization is performed with nloptr (algorithm: NLOPT-LD-LBFGS). We provide an analytical gradient. The Threshold for the outer loop is 1e-9 while the threshold for the inner loop is 1e-13. We use squarem and a C++ implementation for the computation of the market shares to speed up the contraction. We also parallelize the contraction over markets using 7 independent cores. Now we formally describe the instruments included in each test.

Instruments

- J(1): differentiation instruments + exogenous characteristics (polynomial terms) + cost shifters (15 instruments/ degrees of over-identification:8)

- I(1): first stage instruments: instruments J(1). testing instruments: Interval Instruments: 7 instruments. Points chosen as follows: $\{\hat{\mu}, (\hat{\mu} + k(max(0.25, \hat{\sigma})), k(max(0.25, \hat{\sigma}))\}$ $(for\ k = 1, 2, 3)$

- J(2): first stage: instruments: instruments J(1). second stage instruments: optimal instruments (approximation of $\mathbb{E}\left[\frac{\partial \rho_j^{-1}(s_t, x_{2t}, \lambda)}{\partial \lambda} \middle| z_t\right]$)+ exogenous characteristics (polynomial terms) + cost shifters (12 instruments)

- I(2): first stage instruments: instruments J(2). Testing instruments: Interval Instruments: 7 instruments. Points chosen as follows: $\{\hat{\mu}, (\hat{\mu} + k(max(0.25, \hat{\sigma})), k(max(0.25, \hat{\sigma}))\}$ $(for\ k = 1, 2, 3)$

Power against local alternatives We now assess the local power properties of our test by assuming that the random coefficient v_i is distributed according to a local alternative. Namely, we assume $v_i \sim \left(1 - \frac{1}{\sqrt{T}}\right)\mathcal{N}(2, 1) + \frac{1}{\sqrt{T}}Y$ where Y is an alternative distribution including exponential, Chi-square, Student, Uniform. We ensure that Y has mean 2 and variance 1. The results are reported in 13. First, we can observe that except for the uniform local alternative, our test appears to have non-trivial power against all the other local alternatives. For the exponential and chi-square distributions, it is clear that our test with interval instruments outperforms the Sargan-J test with traditional instruments. For the student local alternative, the results seem quite unstable for small sample sizes but as T increases, interval instruments also seem to perform better. For the uniform alternative, it appears that we don't have power against this local alternative.

Table 13: Empirical power, local alternatives (1000 replications)

Number of markets	T=50				T=100				T=200			
Test type	J	I	J	I Local	J	I Local	J	I	J	I Local	J	I Local
Exponential	0.266	0.704	0.227	0.677	0.222	0.869	0.272	0.868	0.236	0.982	0.394	0.975
Chi-square	0.217	0.219	0.134	0.174	0.13	0.167	0.096	0.151	0.099	0.171	0.086	0.15
Student	0.212	0.139	0.33	0.436	0.115	0.115	0.127	0.093	0.082	0.13	0.134	0.312
Uniform	0.198	0.1	0.126	0.074	0.107	0.062	0.095	0.051	0.073	0.049	0.084	0.044

D.3 Finite sample performance of Interval instruments for estimation

Practical implementation of the estimation procedure To assess the performance of our instruments in estimating the non-linear parameters with a flexible distribution of random coefficients, we simulate data with a distribution of random coefficients following a mixture of gaussians and we estimate the parameters of this mixture. For each setting, we estimate the model for 1050 replications. We select the replications with an objective function below a certain threshold (in order to avoid local minima). Minimization is performed with nloptr (algorithm: NLOPT-LD-LBFGS). We provide an analytical gradient, which we describe subsequently. The Threshold for the outer loop is 1e-9 while the threshold for the inner loop is 1e-13. We use squarem and a C++ implementation for the computation of the market shares to speed up the contraction. We also parallelize the contraction over markets using 7 independent core. Before we formally define the different sets of instruments, let us present the estimation procedure when the distribution of random coefficients is assumed to be a mixture.

Instruments Now we formally describe the instruments present in each different sets used for estimation

- Differentiation instruments: differentiation instruments + exogenous characteristics (polynomial terms) + cost shifters (20 instruments)

- Optimal instruments are computed in two stages. The first stage instruments consist of differentiation instruments and exogenous characteristics (polynomial terms). Second stage instruments consist of polynomial terms of exogenous characteristics and the approximation of optimal instruments proposed in Reynaert and Verboven (2014) (approximation of $\mathbb{E}\left[\frac{\partial \rho_j^{-1}(s_t, x_{2t}, \lambda)}{\partial \lambda} \middle| z_t\right]$). The set called optimal instruments includes 15 instruments.

- Interval Instruments are computed in two stages. The first stage instruments consist of differentiation instruments and exogenous characteristics (polynomial terms). Second stage instruments are the interval instruments couples with some exogenous characteristics. A total of 23 instruments. The points in the support to compute the interval instruments are chose as follows: we take equally spaced points in the interval $\{\beta_{3L} - 0.5(\beta_{3H} - \beta_{3L}), \beta_{3H} + 0.5(\beta_{3H} - \beta_{3L})\}$.

Comparison of the performance between the different sets of instruments We now report the mean biases and the empirical $\sqrt{MSE}$ of the estimates for each set of instruments and for different sample sizes. We also plot the distributions of estimates for the non-linear

parameters for the different sets of instruments. First, we plot the distribution of estimates obtained when the set of differentiation instruments from Gandhi and Houde (2019) is used with a sample of $T = 200$ markets and $J = 12$ products. We observe that despite a relatively large sample, the differentiation instruments perform rather poorly in estimating the non-linear parameters associated with the mixture of Gaussians. In particular, the estimates of the standard deviation parameters associated to each component are very dispersed and a large portion of the estimates are bunched at zero. Second, we plot the distribution of non-linear estimates obtained with the optimal instruments from Reynaert and Verboven (2014). They tend to perform better than the differentiation instruments as we can see that the estimates are more concentrated around the true value. Yet, it is important to emphasize that the optimal instruments display large failure rates caused by perfect colinearity of the instruments. We report the percentage of replications that subject to perfect colinearity issues for each sample size $(39\%, 34\%, 31\%, 26\%, 23\%)$. Finally, we plot the distribution of estimates for the non linear parameters when we use the interval instruments developed in section **??**. It appears clearly that the interval instruments yield a more concentrated distribution of estimates than the two other sets of instruments. For the sake of conciseness, we do not report the results with a mixture with 3 components but the observations we make with two components are even more exacerbated.

Table 14: Estimation mixture with "differentiation" instruments (1000 replications)

	Parameter	β_0	α	β_1	β_2	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L
Sample size	true	2	-2	1.5	1	-2	0.5	4	0.5	0.25
T=50, J=12	bias	-0.12	0.022	-0.016	-0.018	0.214	0.184	-0.022	-0.045	0.027
	$\sqrt{MSE}$	0.308	0.06	0.215	0.215	0.633	0.734	0.281	0.35	0.075
T=50, J=20	bias	-0.064	0.011	-0.01	-0.011	0.189	0.347	0.022	-0.081	0.025
	$\sqrt{MSE}$	0.231	0.044	0.165	0.166	0.566	0.887	0.184	0.291	0.059
T=100, J=12	bias	-0.058	0.01	-0.012	-0.012	0.233	0.226	0.02	-0.066	0.027
	$\sqrt{MSE}$	0.204	0.041	0.147	0.148	0.592	0.703	0.256	0.305	0.072
T=100, J=20	bias	-0.04	0.006	-0.007	-0.007	0.198	0.423	0.047	-0.101	0.025
	$\sqrt{MSE}$	0.165	0.032	0.117	0.116	0.552	0.89	0.164	0.27	0.055
T=200, J=12	bias	-0.038	0.007	-0.003	-0.003	0.184	0.167	0.011	-0.049	0.019
	$\sqrt{MSE}$	0.152	0.03	0.11	0.11	0.466	0.601	0.176	0.262	0.053

Table 15: Estimation mixture with "Optimal" instruments(1000 replications)

	Parameter	β_0	α	β_1	β_2	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L
Sample size	true	2	-2	1.5	1	-2	0.5	4	0.5	0.25
T=50, J=12	bias	-0.09	0.016	-0.012	-0.013	0.076	0.059	0.026	-0.111	0.01
	$\sqrt{MSE}$	0.296	0.057	0.234	0.232	0.361	0.483	0.212	0.281	0.036
T=50, J=20	bias	-0.046	0.007	0	0.001	0.074	0.11	0.028	-0.089	0.01
	$\sqrt{MSE}$	0.225	0.044	0.178	0.176	0.328	0.563	0.163	0.228	0.033
T=100, J=12	bias	-0.041	0.007	-0.004	-0.003	0.054	0.037	0.019	-0.066	0.007
	$\sqrt{MSE}$	0.202	0.039	0.157	0.158	0.279	0.4	0.154	0.211	0.028
T=100, J=20	bias	-0.029	0.004	-0.003	-0.003	0.074	0.107	0.033	-0.074	0.01
	$\sqrt{MSE}$	0.153	0.03	0.126	0.124	0.311	0.52	0.129	0.194	0.034
T=200, J=12	bias	-0.029	0.005	-0.001	-0.001	0.026	0.011	0.021	-0.061	0.004
	$\sqrt{MSE}$	0.136	0.026	0.111	0.111	0.184	0.313	0.113	0.172	0.018

Table 16: Estimation mixture with Global Interval instruments(1000 replications)

	Parameter	β_0	α	β_1	β_2	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L
Sample size	true	2	-2	1.5	1	-2	0.5	4	0.5	0.25
T=50, J=12	bias	-0.154	0.029	-0.043	-0.045	0.017	0	-0.045	0.004	0.005
	$\sqrt{MSE}$	0.341	0.067	0.257	0.258	0.277	0.391	0.227	0.259	0.024
T=50, J=20	bias	-0.092	0.017	-0.02	-0.021	0.013	0.042	-0.018	-0.003	0.004
	$\sqrt{MSE}$	0.245	0.048	0.19	0.19	0.248	0.415	0.166	0.22	0.021
T=100, J=12	bias	-0.07	0.013	-0.017	-0.019	0.004	-0.012	-0.027	0.005	0.002
	$\sqrt{MSE}$	0.2	0.039	0.161	0.161	0.167	0.282	0.157	0.201	0.013
T=100, J=20	bias	-0.047	0.008	-0.006	-0.007	-0.009	-0.005	-0.008	-0.009	0.001
	$\sqrt{MSE}$	0.158	0.031	0.13	0.129	0.115	0.264	0.115	0.169	0.005
T=200, J=12	bias	-0.039	0.007	-0.004	-0.003	-0.006	-0.027	-0.015	-0.001	0.001
	$\sqrt{MSE}$	0.141	0.027	0.109	0.109	0.088	0.219	0.108	0.164	0.003

Table 17: Estimation mixture with Local Interval instruments(1000 replications)

	Parameter	β_0	α	β_1	β_2	β_{3L}	σ_{3L}	β_{3H}	σ_{3H}	p_L
Sample size	true	2	-2	1.5	1	-2	0.5	4	0.5	0.25
T=50, J=12	bias	-0.134	0.025	-0.023	-0.024	-0.006	-0.005	-0.039	-0.001	0.003
	$\sqrt{MSE}$	0.307	0.059	0.26	0.259	0.251	0.34	0.214	0.244	0.019
T=50, J=12	bias	-0.084	0.016	-0.024	-0.025	0.019	0.033	-0.023	0.01	0.003
	$\sqrt{MSE}$	0.245	0.047	0.188	0.186	0.228	0.38	0.15	0.184	0.018
T=50, J=12	bias	-0.075	0.015	-0.018	-0.016	0	0	-0.028	0.007	0.001
	$\sqrt{MSE}$	0.199	0.039	0.159	0.16	0.127	0.225	0.143	0.164	0.005
T=50, J=12	bias	-0.039	0.007	-0.011	-0.011	-0.003	0.004	-0.01	0.004	0.001
	$\sqrt{MSE}$	0.162	0.032	0.129	0.129	0.104	0.226	0.103	0.125	0.004
T=50, J=12	bias	-0.037	0.007	-0.008	-0.007	0.002	-0.007	-0.016	0.006	0.001
	$\sqrt{MSE}$	0.136	0.026	0.11	0.109	0.091	0.174	0.099	0.123	0.003

Figure 10: Distribution of estimates for non-linear parameters with "Differentiation" instruments ($T = 200, J = 12$)

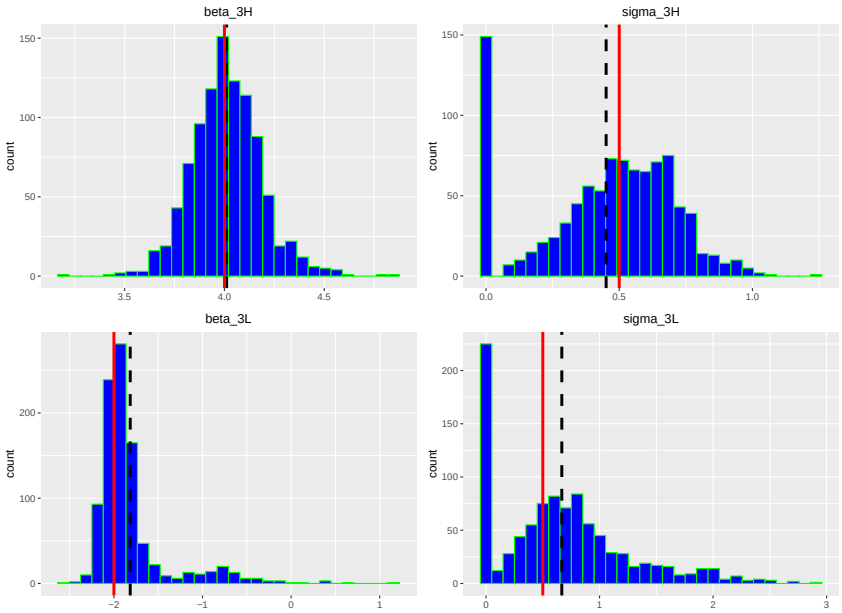


Figure 11: Distribution of estimates for non-linear parameters with "Optimal" instruments ($T = 200, J = 12$)

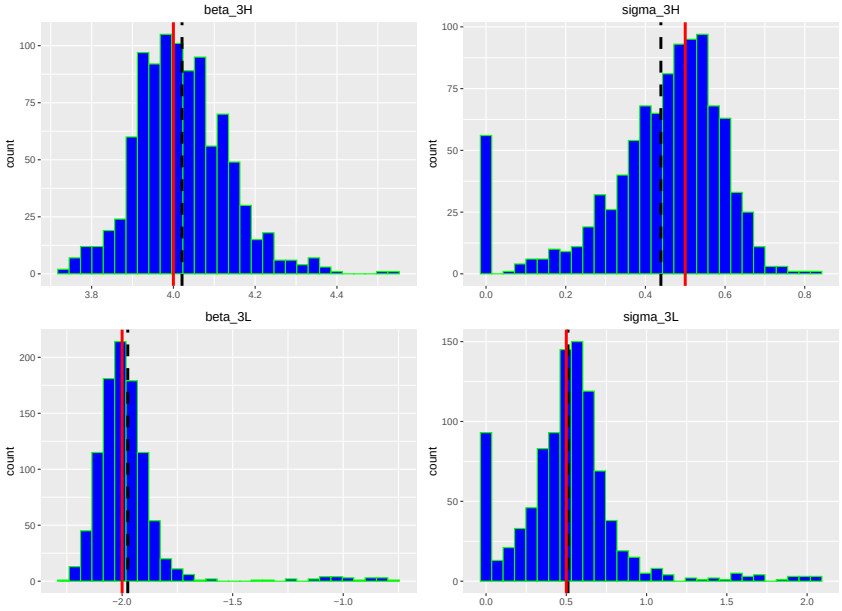


Figure 12: Distribution of estimates for non-linear parameters with "Global Interval" instruments ($T = 200, J = 12$)

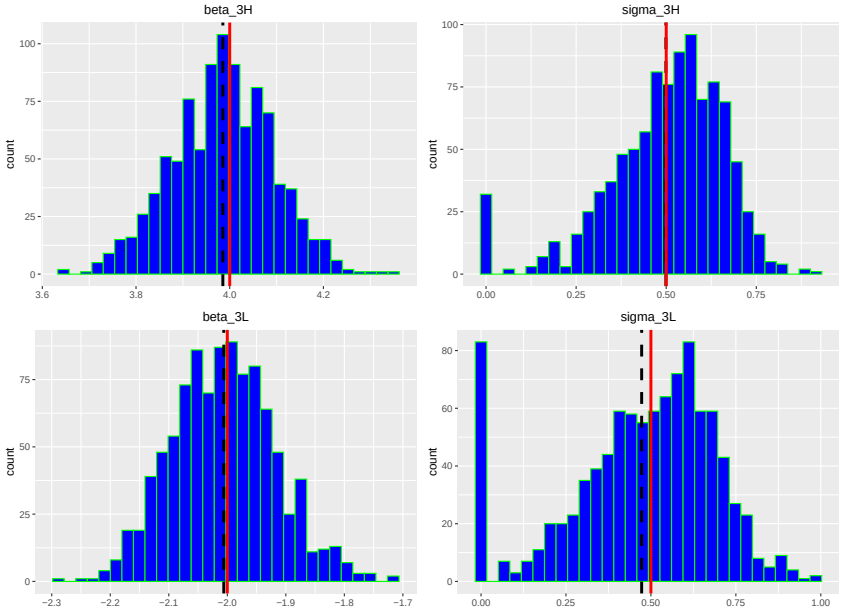
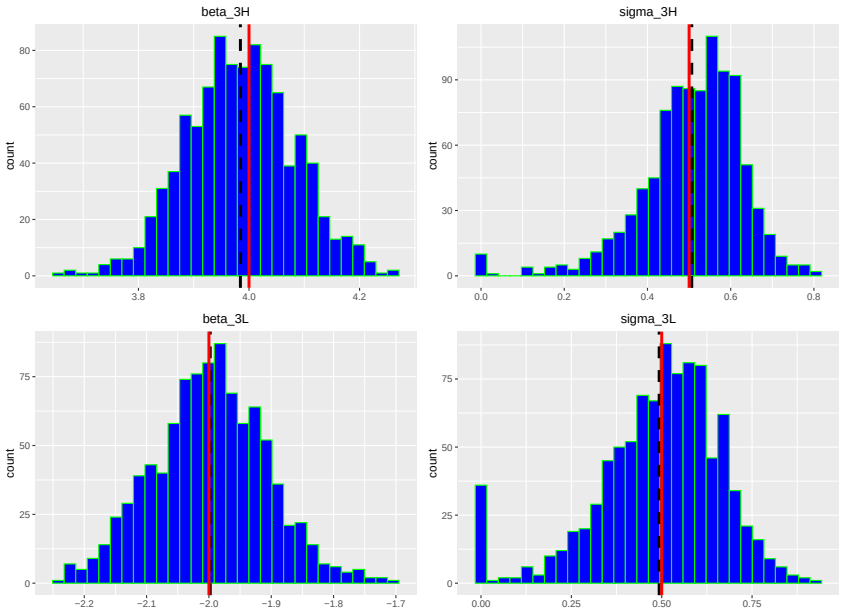


Figure 13: Distribution of estimates for non-linear parameters with "Local interval" instruments ($T = 200, J = 12$)



D.3.1 Estimation with a single Gaussian

Table 18: Estimation with a single Gaussian (1000 replications)

	Instruments	Differentiation						"Optimal"						Interval Global						Interval Local					
	Parameter	β_0	α	β_1	β_2	β_3	σ_3	β_0	α	β_1	β_2	β_3	σ_3	β_0	α	β_1	β_2	β_3	σ_3	β_0	α	β_1	β_2	β_3	σ_3
Sample size	true	2	-2	1.5	1	1.5	0.5	2	-2	1.5	1	1.5	0.5	2	-2	1.5	1	1.5	0.5	2	-2	1.5	1	1.5	0.5
T=50, J=12	bias	-0.16	0.032	-0.031	-0.028	-0.032	-0.004	-0.09	0.018	-0.016	-0.014	-0.018	-0.003	-0.15	0.03	-0.028	-0.026	-0.03	-0.004	-0.15	0.03	-0.028	-0.026	-0.03	-0.001
	$\sqrt{MSE}$	0.292	0.057	0.212	0.209	0.138	0.069	0.27	0.053	0.214	0.211	0.138	0.067	0.288	0.056	0.212	0.209	0.138	0.066	0.286	0.056	0.212	0.209	0.138	0.064
T=50, J=20	bias	-0.091	0.018	-0.022	-0.022	-0.015	0.001	-0.047	0.009	-0.013	-0.013	-0.006	0.001	-0.084	0.017	-0.021	-0.021	-0.013	0	-0.086	0.017	-0.021	-0.021	-0.014	0.002
	$\sqrt{MSE}$	0.209	0.041	0.159	0.16	0.106	0.05	0.199	0.039	0.16	0.161	0.106	0.05	0.206	0.041	0.16	0.16	0.106	0.052	0.208	0.041	0.159	0.16	0.106	0.052
T=100, J=12	bias	-0.088	0.017	-0.001	0	-0.027	0.001	-0.052	0.01	0.007	0.007	-0.02	0.001	-0.082	0.016	0	0.001	-0.026	0.001	-0.074	0.014	-0.016	-0.016	-0.013	0.001
	$\sqrt{MSE}$	0.199	0.039	0.146	0.145	0.1	0.045	0.189	0.037	0.148	0.147	0.099	0.047	0.197	0.039	0.146	0.146	0.1	0.044	0.185	0.036	0.151	0.152	0.099	0.044
T=100, J=20	bias	-0.043	0.009	-0.011	-0.012	-0.006	-0.001	-0.021	0.004	-0.007	-0.008	-0.002	-0.001	-0.04	0.008	-0.011	-0.012	-0.006	-0.001	-0.035	0.007	-0.01	-0.009	-0.004	0
	$\sqrt{MSE}$	0.145	0.028	0.115	0.114	0.075	0.035	0.141	0.028	0.115	0.114	0.075	0.035	0.145	0.028	0.115	0.114	0.076	0.035	0.14	0.027	0.116	0.115	0.076	0.035
T=100, J=20	bias	-0.038	0.007	-0.012	-0.012	-0.004	0.001	-0.017	0.003	-0.006	-0.007	-0.001	0	-0.032	0.006	-0.009	-0.01	-0.004	0	-0.033	0.006	-0.009	-0.01	-0.004	0.001
	$\sqrt{MSE}$	0.132	0.026	0.11	0.11	0.073	0.032	0.127	0.025	0.109	0.109	0.069	0.032	0.129	0.026	0.109	0.109	0.069	0.032	0.129	0.026	0.109	0.109	0.069	0.031

E Empirical application

First stage regression: instruments on price

Table 19: Estimation results - Logit and Nested Logit

	OLS		*instrumental variable*		
	(1)	(2)	(3)	(4)	(5)
Price/income	−0.354***	−2.907***	−2.356***	−2.729***	−2.615***
	(0.041)	(0.133)	(0.124)	(0.053)	(0.052)
log(within market shares)				0.420***	0.407***
				(0.006)	(0.006)
Fuel Cost	−0.210***	−0.138***	−0.247***	−0.074***	−0.126***
	(0.008)	(0.006)	(0.009)	(0.004)	(0.006)
Size(m^2)	0.031	0.001	0.158***	−0.001	0.104***
	(0.038)	(0.040)	(0.041)	(0.025)	(0.026)
Horsepower(KW/100)	0.136	3.151***	2.511***	2.586***	2.431***
	(0.089)	(0.183)	(0.172)	(0.080)	(0.078)
Foreign	0.351***	0.083	0.120*	−0.106**	−0.101**
	(0.064)	(0.073)	(0.070)	(0.046)	(0.044)
Height(m)	0.870***	1.505***	3.487***	1.121***	2.270***
	(0.216)	(0.197)	(0.228)	(0.125)	(0.145)
Gasoline	1.399***	0.625***	1.118***	0.190***	0.422***
	(0.055)	(0.061)	(0.063)	(0.039)	(0.041)
Fuel cost*imes* income	0.020***	−0.002**	0.014***	−0.002***	0.007***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Size *imes* income	−0.005***	−0.002***	−0.006***	0.0003	−0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Horsepower*imes* income	0.009***	−0.026***	−0.017***	−0.027***	−0.024***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Horsepower*imes* time	−0.084***	−0.068***	−0.083***	−0.038***	−0.045***
	(0.006)	(0.007)	(0.007)	(0.004)	(0.004)
Foreign *imes* income	−0.019***	−0.015***	−0.016***	−0.008***	−0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Height *imes* income	−0.006	0.032***	−0.002	0.016***	−0.003
	(0.004)	(0.004)	(0.005)	(0.003)	(0.003)
Height *imes* density	−0.037***	−0.003***	−0.037***	−0.001***	−0.021***
	(0.004)	(0.0003)	(0.004)	(0.0002)	(0.003)
Gasoline*imes* income	−0.016***	−0.003***	−0.010***	0.0004	−0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
X2p2015s		−0.024		−0.019	
		(0.012)		(0.012)	
Constant	−7.937***	−12.482***	−11.171***	−9.144***	−8.506***
	(0.167)	(0.149)	(0.167)	(0.092)	(0.102)
State FE/ Year FE	*Yes*	*No*	*Yes*	*No*	*Yes*
Observations	39,888	39,888	39,888	39,888	39,888

Baseline specifications: logit and nested logit

Construction of the interval instruments

- **Discretization of the support**
- **normalization of the instruments**

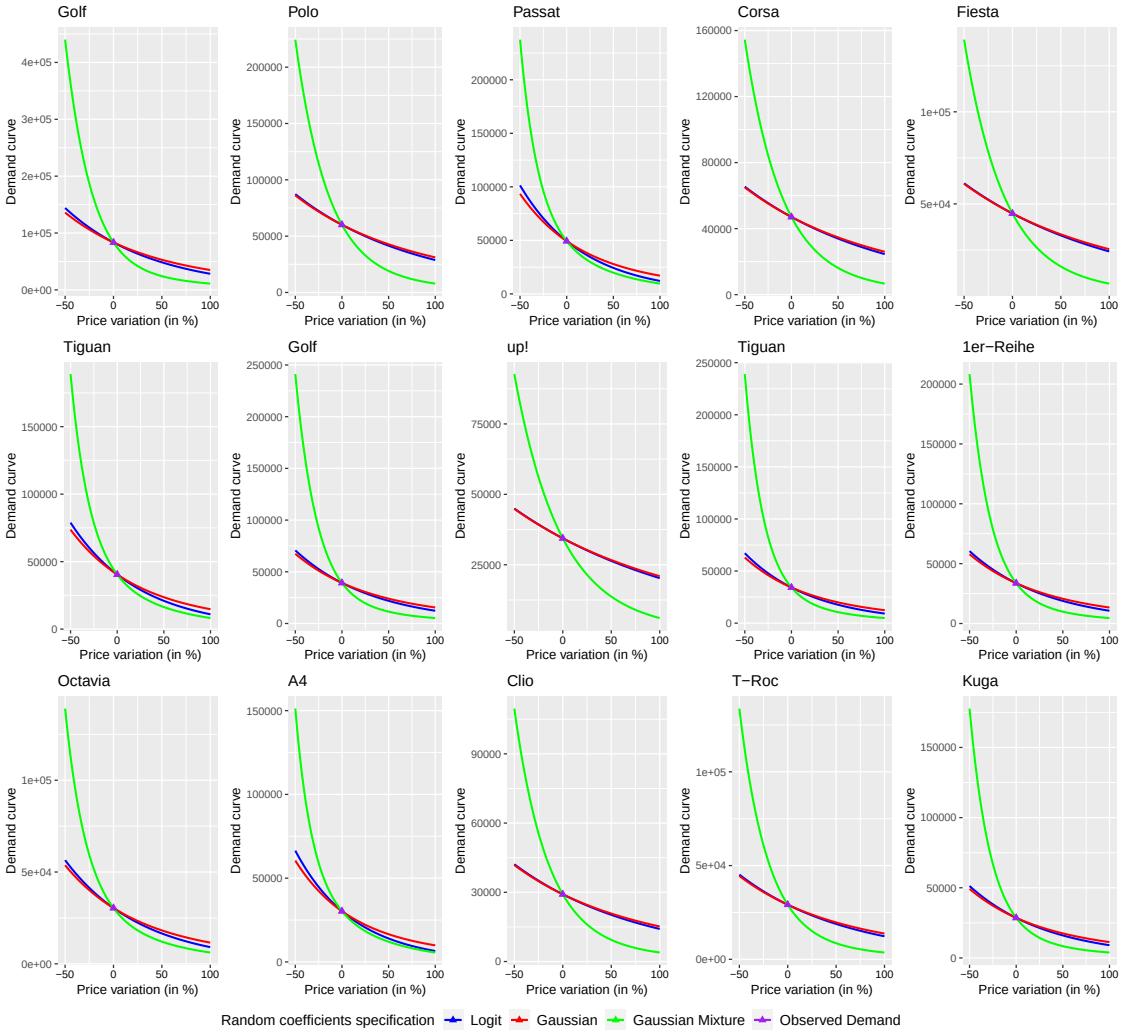
E.0.1 Results differentiation instruments

Table 20: counterfactual quantities under different specifications on RCs (20 most popular cars)

Counterfactual quantity		Price elasticity			Curvature			Marginal cost			Mark-up			Pass-through		
car	Manufacturer	Logit	Gaussian	Mixture	Logit	Gaussian	Mixture	Logit	Gaussian	Mixture	Logit	Gaussian	Mixture	Logit	Gaussian	Mixture
Golf	Volkswagen	-1.09	-0.95	-3.03	1.00	1.14	1.21	1260	-9670	15436	24098	35028	9922	0.92	-	1.30
Polo	Volkswagen	-0.74	-0.70	-2.50	1.00	1.15	1.09	-6643	-14366	9073	23819	31542	8103	1.05	-	1.09
Passat	Volkswagen	-1.43	-1.21	-2.27	1.00	1.17	1.57	9488	-1033	17826	24631	35153	16294	1.02	-	2.65
Corsa	PSA	-0.66	-0.63	-2.28	1.00	1.14	1.07	-8432	-11246	8410	24088	26902	7246	1.02	-	1.12
Fiesta	Ford	-0.62	-0.60	-2.18	1.00	1.15	1.07	-8983	-10806	7657	23487	25310	6847	1.03	-	1.10
Tiguan	Volkswagen	-1.32	-1.14	-2.28	1.00	1.17	1.55	6831	-2919	16211	24118	33868	14738	1.01	-	2.62
Golf	Volkswagen	-1.17	-1.03	-3.12	1.00	1.18	1.27	3128	-7932	16582	23828	34888	10374	0.99	-	1.41
up!	Volkswagen	-0.53	-0.52	-1.92	1.00	1.14	1.05	-11231	-17703	4594	23278	29749	7453	1.04	-	0.96
Tiguan	Volkswagen	-1.34	-1.15	-3.09	1.00	1.19	1.38	7051	-4117	19186	23842	35009	11706	1.01	-	1.66
1er-Reihe	BMW	-1.16	-1.03	-3.09	1.00	1.18	1.28	3845	-769	19179	25138	29753	9805	0.99	-	1.39
Octavia	Volkswagen	-1.23	-1.08	-2.33	1.00	1.17	1.50	4629	-4504	15464	24211	33345	13377	1.01	-	2.34
A4	Volkswagen	-1.56	-1.30	-2.26	1.00	1.19	1.56	13209	1995	20260	25865	37079	18814	1.01	-	2.66
Clio	Renault	-0.73	-0.70	-2.49	1.00	1.16	1.10	-6240	-8684	9817	23120	25563	7063	1.03	-	1.17
T-Roc	Volkswagen	-0.87	-0.81	-2.80	1.00	1.17	1.14	-3645	-12275	11578	23798	32427	8575	1.06	-	1.16
Kuga	Ford	-1.16	-1.03	-3.09	1.00	1.18	1.28	3654	-518	18214	23684	27856	9124	1.03	-	1.39
Golf	Volkswagen	-1.10	-0.99	-2.34	1.00	1.16	1.44	1548	-7284	13678	23929	32762	11799	0.96	-	2.13
A-Klasse	Daimler	-1.28	-1.10	-3.07	1.00	1.19	1.35	6608	562	20662	25066	31112	11013	1.01	-	1.56
Golf	Volkswagen	-1.05	-0.94	-2.33	1.00	1.16	1.42	417	-8115	13135	24177	32710	11460	0.72	-	2.11
Golf	Volkswagen	-1.18	-1.05	-3.15	1.00	1.18	1.27	3202	-8230	16705	23921	35353	10418	0.98	-	1.40
Octavia	Volkswagen	-1.05	-0.95	-3.02	1.00	1.17	1.21	380	-8835	14808	23862	33077	9433	0.78	-	1.30

Counterfactual quantities under different specifications

Figure 14: Estimated demand functions under different specifications



Chapter 3: Selecting Strong and Exogenous Instruments via Structural Error Criteria

Abstract

Instrumental variables (IVs) allow consistent estimation of the causal effect of endogenous variables on outcomes. However, if IVs are not exogenous and jointly strong, estimators are inconsistent and t-test based Gaussian confidence intervals are invalid. Thus, in this paper I design a procedure to select a subset of strong and exogenous IVs among a larger set of potentially weak and / or endogenous IVs in a linear setting. To do so I formally build losses, risks and risk estimators which are based on the structural errors being implicitly minimized when performing IV estimation. I shed light into the empirical and theoretical properties of the risks and find that IV subset selection via risk estimators minimization consistently select strong and exogenous subsets of IVs for the two-stage least squares (2SLS) estimator. More specifically, efficiency and consistency results are established by considering standard asymptotics, weak IV asymptotics and locally invalid IV asymptotics, while maintaining the total number of IVs fixed. I confirm the performances of my IV selection procedures against competing ones' using Monte Carlo simulations and lastly I estimate the causal effect of pre-trial detention on offenders guilt by selecting judge dummy IVs in the first stage.

Keywords: Instrument Selection, Valid Instrument, Weak Instrument, Model Selection
JEL Codes: C52, C14

1 Introduction

IVs are used to estimate and infer on the causal effect of endogenous variables on outcomes. Yet applied researchers still struggle in their choice of IVs and specification because of a complicated trade-off between the quality and number of IVs, and the bias, the efficiency, and the asymptotic distribution of their IV estimator. Thus, when econometricians have multiple IVs at their disposal one conventional solution is to use all of them. But for this solution to work all IVs must be exogenous and jointly strong, which may not be the case in practice. Another common solution is to use a single IV, typically the IV for which exogeneity can be best justified. However, this choice is not data-driven and can actually be quite arbitrary. Instead, in this paper I assume that there exists a strong and exogenous subset of IVs and propose data-driven methods to find this subset based on out-of-sample validation. To do so I formally define losses, risks and propose risk estimators in order to consistently choose strong and exogenous IV subsets even in the presence of endogenous and weak IVs.

Prior work It is well-known that in the presence of weak or endogenous and possibly many IVs the traditional 2 stage least squares estimator (2SLS) is inconsistent and confidence intervals (CI) built from its Gaussian asymptotics have low coverage. See Stock, Wright, and Yogo (2002), Hahn and Hausman (2003), Hahn, Hausman, and Kuersteiner (2004), Kiviet and Kripfganz (2021) for general reviews on weak, many weak, many, and endogenous IVs problems respectively. The literature has treated each problem separately even tough in practice they are likely to occur at the same time.

Regarding weak IVs the literature has developed in the last 25 years with the maintained assumption that IVs are exogenous. It has mainly focused on detecting weak IVs, see Stock and Yogo (2005), Kleibergen (2007), and Olea and Pflueger (2013), and on inference procedures which are robust to weak identification, see Anderson and Rubin (1949), Kleibergen (2002), Moreira (2003), with their subvector counterparts, see Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), and their nonlinear first stage counterparts, see Antoine and Lavergne (2022) and Boucher (2022). The current consensus being that if weak IVs are detected such inference procedures should be used.

The literature on many (weak) IVs has studied the behavior of different k-class estimators under a many exogenous IVs or a many weak exogenous IVs assumption. Compared to 2SLS, other k-class estimators reduce finite sample bias, can be consistent under specific assumptions on the types of asymptotics, see Hahn et al. (2004), and can allow for valid inference,

see Mikusheva and Sun (2021) and Andrews, Marmer, and Yu (2019) for reviews on inference under many weak IVs, and many IVs asymptotics respectively. Part of this literature has focused on regularizing these estimators, selecting IVs based on first stage fit, see Donald and Newey (2001), Bai and Ng (2010), Carrasco (2012), Belloni, Chen, Chernozhukov, and Hansen (2012), Chen, Chen, and Lewis (2021).

Lastly, to deal with (many) endogenous but strong IVs, the literature has focused on detecting endogeneity at the vector or subvector level using overidentifying restrictions types of tests with more recent papers allowing for heteroskedasticity and possibly many IVs, see Sargan (1958), K. Newey (1985), Hahn and Hausman (2002), Carrasco and Doukali (2021), and also focused on devising procedures to select the exogenous IVs directly or indirectly via regularization, see Andrews (1999), Hall and Peixe (2003), Caner (2009), Kang, Zhang, Cai, and Small (2016), Windmeijer, Farbmacher, Davies, and Smith (2018), Gautier and Rose (2021). More recently, several papers on sensitivity analysis derive falsification sets[1] for the linear IV model with homogenous effect and potentially endogenous IVs, see Masten and Poirier (2021), and Apfel and Windmeijer (2022).

Contribution In a context with a finite number of IVs which is most common in practice, a strong argument can be made in favor of constructing criteria in order to select a strong and exogenous subset of IVs. Such criteria do not yet exist and would allow to remove the irrelevant and misleading information contained in the full set of IVs. Indeed, in case all the IVs are exogenous or exogenous and strong, IV selection can only improve inference and lower finite sample bias by reducing the number of IVs and improving their overall strength. On the other hand when IVs can be endogenous selection is a necessity otherwise the true causal effect cannot be estimated consistently and valid inference cannot be performed. But thus far current popular methods for IV selection are either unable to pick exogenous IVs if endogenous IVs are present as in Donald and Newey (2001), Bai and Ng (2010), Belloni et al. (2012), or Carrasco and Tchuente (2016), either unable to pick up strong exogenous IVs if weak exogenous IVs are present as in Andrews (1999) or Kang et al. (2016), either require an a priori consistent estimator as in Donald and Newey (2001) or Windmeijer et al. (2018).

Accordingly, I consider a linear IV model where the total number of IVs remain fixed, where IVs can be correlated but also individually weak or strong, and where some IVs enter the structural equation. This setting with a finite number of IVs and an unknown subset of

[1]Falsification sets contain the effects which are estimated by different 2SLS estimators when using a single IV for instrumentation and the rest of the IVs as control variables.

exogenous IVs is the most common in practice. Note that the endogenous IVs may directly affect the outcome or indirectly through some unobserved regressor which they are correlated to. The goal is then to find a subset of strong and exogenous IVs among the full set of IVs. This is a model selection problem (see Arlot and Celisse (2010) and Bates, Hastie, and Tibshirani (2021) for recent surveys on model selection) thus I design selection criteria for IV subsets based on the structural error being implicitly minimized in an IV setting. Hence I define three prediction losses for IV subsets: A loss based on the exogeneity condition; The mean squared error of prediction where the endogenous variable has been projected on the IVs; And the mean squared error of prediction. Then I define the corresponding risks, which are average prediction losses, and their corresponding cross-validation estimators for IV subsets. Finally, the IV subset which minimizes the risk is selected and used as IVs whereas the rest of the IVs are considered endogenous and therefore used as control variables. In terms of theory, I provide a decomposition of the risks and show that the IV subset selection procedures are efficient and consistent. If in the full set of IVs there exists a subset which is strong and exogenous, it will be selected with probability one at the limit. These results are established by allowing for weak IVs (in the sense of Staiger and Stock (1997) and Andrews and Cheng (2012)), and by allowing for locally invalid IVs (as in the literature on local misspecificaiton, see Maasoumi and Phillips (1982), and on sensitivity analysis, see Andrews, Gentzkow, and Shapiro (2017)). I confirm these findings by looking at the performances of the 2SLS estimator and at the performances of the weak-identification robust inference procedure from Moreira (2003) in an extensive simulation exercise. I also apply my methods and select judge dummy variables in order to estimate the effect of pre-trial detention on the likelihood of being found guilty.

Outline The outline of the paper is as follows. In the second section of this paper I present the linear IV model when considering subset of IVs and interpret the estimated parameter in terms of losses. Then in the third section I define the losses, risks and risk estimators which are relevant for IV subset selection and present the selection procedure. In the fourth section I present theoretical guaranties that strong and exogenous sets of IVs are systematically selected compared to weak and endogenous subsets. In the fifth section I show through simulations that the methods consistently select strong and exogenous IV subsets and therefore yield estimators and inference procedures with great performances. In the sixth section I apply my methods and estimate the effect of pre-trial detention on offenders' probability of being found guilty. I conclude in the seventh and final section.

2 Model, estimator, and loss interpretation

2.1 Linear IV model with endogenous IVs

Consider the following linear IV model with outcome y_i, a single endogenous variable x_i, K_z IVs z_i among which z_{iE} is a strong exogenous subset and its complement $z_{i\overline{E}}$ is endogenous and enter the structural equation linearly

$$y_i = x_i\beta + z'_{i\overline{E}}\alpha + u_i, \qquad \mathbb{E}(u_i|z_i) = 0, \qquad \mathbb{E}(u_i^2|z_i) = \sigma_u^2, \qquad \mathbb{E}(u_i v_i|z_i) = \rho \tag{2.1}$$

$$x_i = z_i'\pi + v_i, \qquad \mathbb{E}(v_i z_i) = 0, \qquad \mathbb{E}(v_i^2|z_i) = \sigma_v^2 \tag{2.2}$$

for $i = 1, \dots, n$ where $z_i'\pi \equiv z'_{iE}\pi_E + z'_{i\overline{E}}\pi_{\overline{E}}$ and π_E is non-zero and fixed with n. To simplify exposition and calculation, I also impose that the data is centered. The model characterized by the structural equation or second stage (2.2) and the reduced form equation or first stage (2.1) is very common in applied work. A single causal effect β is of interest and a few IVs are used to try to consistently estimate it, see the example below. The set of IVs may be the result of some interactions between IVs and exogenous controls, or the result of modeling non-linearly the relation between the endogenous variable and IVs. The only departure from the usual linear IV model is that only the subset of IVs z_{iE} is exogenous. Intuitively, the subset of IVs $z_{i\overline{E}}$ does not satisfy the exclusion restriction and affects y_i either directly or indirectly through some unobservable, see figure 1 in appendix A.1 for some visualization. To go further, endogeneity as in correlation between u_i and z_i as in Masten and Poirier (2021) is already incorporated into this model. Indeed, $z'_{i\bar{E}}\alpha$ can be viewed as the linear projection of the "error" on the IVs[2]. In fact, when $\alpha = 0$ then the model reduces to the usual IV model under exogeneity.

Note that the selection procedures and the formal results are later established under an independent and identically distributed and conditional homoskedasticity assumption which can be relaxed to allow for conditional heteroskedasticity without modifying the selection criteria. Extending the results to the case where x_i is a vector is also possible but requires a

[2]If the model is defined as

$$y_i = x_i'\beta + \tilde{u}_i, \qquad \mathbb{E}(z_i\tilde{u}_i) = \eta, \qquad Var(\tilde{u}_i^2|z_i) = \sigma_u^2, \qquad Cov(\tilde{u}_i, v_i|z_i) = \rho$$

$$x_i = z_i'\pi + v_i, \qquad \mathbb{E}(v_i z_i) = 0, \qquad \mathbb{E}(v_i 2|z_i) = \sigma_v^2$$

then $\tilde{u}_i$ can be projected on z_i so that $\tilde{u}_i = z_i'\zeta + u_i$ where $\mathbb{E}(u_i z_i) = 0$ and $\zeta \equiv \mathbb{E}(z_i z_i')^{-1}\mathbb{E}(z_i\tilde{u}_i) = \mathbb{E}(z_i z_i')^{-1}\eta$. Consequently, $\zeta = 0 \Leftrightarrow \eta = 0$ although if $\exists j : \zeta_j \neq 0$ then $\zeta_{j'} = 0 \not\Leftrightarrow .\eta_{j'} = 0$.

more complex modelization of IV weakness at the vector level and additional assumptions to obtain the consistency of the selection procedures. As for exogenous controls, they can be projected out a la Firsch-Waugh with little consequences and are therefore omitted. A setting with many IVs or a fully non-linear modelization of the first stage are outside the scope of this paper.

Clearly β is identified when using z_{iE} as IVs and $z_{i\overline{E}}$ as control variables but in practice it cannot be estimated consistently because E is unknown. For this reason it is key to reformulate the model and consider the 2SLS estimator for a specific subset of IVs but before that consider the following simple example of the model characterized by (2.1) and (2.2).

Example: weather IVs Consider estimating the following demand curve for fish at the Fulton fish market as in Graddy (2006)

$$Q_i = P_i\beta + X_i'\delta + u_i$$

where i denotes the day, Q_i the total amount of fish sold during day i, P_i the average daily price of the fish, and X_i various control variables. Clearly P_i is endogenous because it is determined simultaneously with Q_i. Consequently, the demand curve must be augmented by the following reduced form first stage equation

$$P_i = \pi_1 cold_i + \pi_2 wind_i + \pi_3 rain_i + \pi_4 stormy_i + \pi_5 mixed_i + X_i'\gamma + v_i$$

where $cold_i$, $wind_i$, $rain_i$, $stormy_i$, and $mixed_i$ are available weather IVs. Then to identify β it must be that the weather variables are cost shifters which affect demand only through price. But some IVs such as $wind_i$ may not affect supply significantly, and some IVs such as $cold_i$ may not have a direct effect on demand. Thus, it is unclear which weather IV is truly exogenous and strong. Therefore, instead of arguing (with difficulty) for the validity of specific weather IVs it seems much more natural to select a strong and exogenous subset of IVs in a data driven way.

2.2 Subset model and subset IV 2SLS

Before describing the IV subset selection method, the model and the 2SLS estimator for a given subset of IVs must be defined and additional notations must be introduced.

In the rest of the paper, let $\mathcal{S}$ denote the collection of all non-empty subsets of

$$\{z_{i1}; z_{i2}; \ldots; z_{iK_z}\}$$

The cardinality of $\mathcal{S}$ is therefore $2^{K_z} - 1$. The complement of S is denoted as $\overline{S}$ in the sense that $S \cup \overline{S} = \{z_{i1}; z_{i2}; \ldots; z_{iK_z}\}$ and $S \cap \overline{S} = \emptyset$. The IVs associated to $S \in \mathcal{S}$ are denoted as z_{iS} which is a random vector of dimension $s = |S|_0$ where $|\cdot|_0$ denotes the counting norm, and π_S is the subvector of π of dimension s associated to S. Let $\Sigma \equiv \mathbb{E}(z_i z_i')$, and for any $S \in \mathcal{S}$ let $\Sigma_S = \mathbb{E}(z_{iS} z_{iS}')$. In addition, to simplify notations, I denote $w_i \equiv (y_i, x_i, z_i')'$ the observed variables for individual i. Furthermore, let $y \equiv (y_1, y_2, \ldots, y_n)'$ be the $n \times 1$ vector of stacked outcomes over the sample, let $x \equiv (x_1, x_2, \ldots, x_n)'$ be the $n \times 1$ vector of stacked endogenous variables, let $z \equiv (z_1 \ z_2 \ \ldots \ z_n)'$ be the $n \times K_z$ matrix of stacked IVs, and let $w \equiv (w_1 \ w_2 \ \ldots \ w_n)'$ be the $n \times (K_z + 2)$ matrix of stacked observed variables. Similarly, $u = (u_1, u_2, \ldots, u_n)$, $v = (v_1, v_2, \ldots, v_n)$, and $z_S = (z_{1S} \ z_{2S} \ \ldots \ z_{nS})'$ for any $S \in \mathcal{S}$.

As discussed the subset of exogenous IVs E is unknown a priori, thus without selection it is not possible to estimate β consistently. As a consequence, some candidate S has to be considered for instrumentation and its complement $\overline{S}$ has to enter as a vector of control variables. When S is picked for instrumentation and $S \subset E$ the model is called valid, whereas when $S = E$ the model is called the oracle model. In both these cases β can be estimated consistently, see figure 2 in appendix A.1. If the full vector of IVs z_i is used for instrumentation, then no instrument should be included as a control variable. Consequently, if subset S is considered for instrumentation the model can be rewritten as

$$y_i = x_i\beta + z_{i\overline{S}}'\alpha_{\overline{S}} + u_{i\overline{S}}, \qquad x_i = z_{iS}'\pi_S + z_{i\overline{S}}'\pi_{\overline{S}} + v_i, \qquad \mathbb{E}(z_i v_i) = 0 \tag{2.3}$$

for some $(\alpha_{\overline{S}}, u_{i\overline{S}})$. Since $z_{i\overline{S}}$ is considered a vector of control variables, it can be projected out a la Frisch-Waugh. Thus, except in the rest of the paper except the proofs, I denote $(y_i, x_i, z_{iS}) \equiv (y_i, x_i, z_{iS}) - BLP((y_i, x_i, z_{iS})|z_{i\overline{S}})$ where $BLP(\cdot|z_{i\overline{S}})$ is the best linear projection on $z_{i\overline{S}}$. Consequently, the model instrumented by S (2.3) can be rewritten as

$$y_i = x_i\beta + u_{iS}, \qquad \mathbb{E}(u_{iS} z_{i\overline{S}}) = \mathbb{E}(x_i z_{i\overline{S}}) = 0 \tag{2.4}$$

$$x_i = z_{iS}'\pi_S + v_i, \qquad \mathbb{E}(v_i z_i) = 0 \tag{2.5}$$

where $u_{iS} = z_{i\overline{E}}'\alpha + u_i$, $Var(u_{iS}|z_i) = \sigma_u^2$, $Var(v_i|z_i) = \sigma_v^2$ and $Cov(u_{iS}, v_i|z_i) = \rho$. Hence, if subset S is considered for instrumentation the fact that β can be estimated consistently depends on whether $\mathbb{E}(z_{iS} u_{iS})$ is close to zero or not, which is the case when α is close to zero and when $\overline{S}$ is close to $\overline{E}$, and on whether π_S is equal to zero or not.

From the model instrumented by S characterized by (2.4) and (2.5) I define the 2SLS subset IV estimator

$$\hat{\beta}_S = \frac{x' P_{z_S} y}{x' P_{z_S} x} = \beta + \frac{x' P_{z_S} u_S}{x' P_{z_S} x}$$

where $P_{z_S} = z_S(z_S'z_S)^{-1}z_S$ is the orthogonal projection on z_S. For exposition, the paper focuses on 2SLS, but the criteria and results developed for 2SLS also work for any k-class estimators (which possess moments). In fact, using k-class estimators allow to relax some of the assumptions required for the consistency of the selection procedures.

2.3 Interpreting the causal effect in terms of losses

Having rewritten the model and estimator when subset S is considered for instrumentation, it is necessary to understand what are the losses of interest in the linear IV model. This will guide the choice of criteria for IV subset selection.

Traditionally, the causal effect β is defined as the parameter for which the exogeneity condition is satisfied. Consequently, when instrumenting with S the structural parameter β is the minimizer of a weighted sum of the squared correlations between the subset of IVs S and the error

$$\beta(S,W) = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z_{iS}')W\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$$

for some symmetric full ranked weighting matrix W. $\beta(S,W)$ is the set of parameters that can be estimated given weights W and IVs S, it is the set of pseudo true values or target set induced by the minimization of the loss based on the exogeneity condition. Taking some empirical counterpart of the population moments in the objective and assuming validity of the IVs will yield the different k-class and GMM estimators. A natural candidate for W is $W = \Sigma_S^{-1}$, ie the correlation structure of the IVs is controlled for. Therefore, given IV set S let β_S be the target parameter set in IV estimation which minimizes the following exogeneity based loss

$$\beta_S = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z_{iS}')\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta})) \tag{2.6}$$

If S is irrelevant as in $\pi_S = 0$ then $\beta_S = \mathbb{R}$, if S is relevant but possibly endogenous as in $\pi_S \neq 0$ then $\beta_S = \beta + (\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha$, if S is relevant and exogenous as in $\pi_S \neq 0$ and $\pi_S'\mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha = 0$ then and only then is the parameter to be estimated equal to the causal effect $\beta_S = \beta$. Hence, as the mean square error is the loss of interest in linear model, a loss of specific interest in linear IV models is based on the exogeneity condition.

More historically, the goal behind IV estimation is trying to perform ordinary least squares minimization while circumventing the endogenous nature of the regressor. So instead of

using x_i as a regressor only its (potentially) exogenous part $BLP(x_i|z_{iS}) = z'_{iS}\pi_S$ the best linear projection of x_i on z_{iS} is used. In that sense β_S can be rewritten as

$$\beta_S = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2)$$

As before $\beta_S = \mathbb{R}$ if $\pi_S = 0$, $\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha$ if $\pi_S \neq 0$, and $\beta_S = \beta$ if $\pi_S \neq 0$ and $\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0$. Hence, a second natural loss to minimize in a linear IV model given IV set S is the mean square error with the endogenous variable projected on the IVs. See appendix D.1 and Proposition 4.1 and 4.2 for formal results.

Having defined the IV subset estimator and the losses of interest in linear IV models for any candidate for instrumentation S, I formally design the criteria for the selection of IV subsets in the next section.

3 Risks for IV sets

In statistics and machine learning model selection is now understood within a common framework, see Arlot and Celisse (2010), and IV subset selection can also be understood within this framework. Thus, in this section I introduce prediction losses, risks, and risk estimators for IV subsets. In practice these risk estimators are computed for each IV subset and the subset which minimizes them will be chosen for instrumentation. Lastly, I describe the desirable properties of IV subset selection procedures.

3.1 Model selection for IV subsets

Prediction losses In model selection the performances of an estimator (or of an IV subset in this case) are measured with prediction losses. Indeed, if the right model is picked then it should perform very well with new data. Prediction losses are usually defined as average prediction errors or average out-of-sample discrepancies with respect to a new observation w^* conditional on the original sample $(w_i)_{i=1}^n$ where w^* has the same DGP but is independent of $(w_i)_{i=1}^n$. Let $\mathbb{E}_n(\cdot) \equiv \mathbb{E}(\cdot|(w_i)_{i=1}^n)$ denote the expectation conditional on the original sample,

then define the following losses for any IV subset S

$$L_{EXO}(w^*; \tilde{\beta}, S) = \mathbb{E}_n \left((y^* - x^* \tilde{\beta}) z_S^{*'} \right) \Sigma_S^{-1} \mathbb{E}_n \left(z_S^* (y^* - x^* \tilde{\beta}) \right)$$
$$L_{PMSE}(w^*; \tilde{\beta}, S) = \mathbb{E}_n \left((y^* - z_S^{*'} \pi_S \tilde{\beta})^2 \right)$$
$$L_{MSE}(w^*; \tilde{\beta}) = \mathbb{E}_n \left((y^* - x^* \tilde{\beta})^2 \right)$$

L_{EXO} corresponds to an out-of-sample counterpart of the exogeneity based loss defined in the previous section, if a new observation w^* is at disposal then the correlation between the error and the IVs should be small and therefore L_{EXO} should be small. L_{PMSE} corresponds to an out-of-sample counterpart of the mean square error of prediction after projecting the endogenous variable on the IVs, again for a new observation it should naturally be small if the right IV subset was picked. L_{MSE} is the mean square error of prediction of the structural equation. While L_{MSE} is not of direct interest in the context of linear-IV models, it is already extensively used for nuisance parameter selection (bandiwdth, Lasso penalty, basis size, etc...) of various IV estimators and procedures, see Chernozhukov, Hansen, and Spindler (2015) or Kang et al. (2016), thus its IV subset selection properties are also studied. As will be shown in the next section, the risk based on the mean square error of prediction can actually select strong and exogenous IV subsets in certain conditions.

Other types of prediction losses can be discussed, but they are less appealing. Excess losses are versions of losses which are "centered" with respect to the true β, but in the linear-IV context "centering" is difficult because β is identified by IV subset E which is unknown a priori. Losses which do not depend on S are in line with the literature on predictors selection in linear models but too different from the losses being implicitly minimized during IV estimation. Alternatively, losses built from the log-likelihood of (y^*, x^*) given z_S^* could be used. However, they require the first stage (2.5) to be causal which is very hard to argue for. Finally, integrated losses are much more suited to a non-linear setting with fewer variables. See Arlot and Celisse (2010) for formal definitions.

Risks To properly assess the performances of $\hat{\beta}_S$, losses have to be evaluated at $\hat{\beta}_S$ and averaged because they are random. These average prediction losses are called risks. The risks

for any IV subset S to consider are therefore

$$R_{EXO}(S) = \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(y^* - x^*\hat{\beta}_S)\right)\right)$$
$$R_{PMSE}(S) = \mathbb{E}\left(\mathbb{E}_n\left((y^* - z_S^{*'}\pi_S\hat{\beta}_S)^2\right)\right)$$
$$R_{MSE}(S) = \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)^2\right)\right)$$

for some new observation w^*. The risks are thoroughly decomposed and interpreted in section 4.1. Two other risks have been formalized in the literature, Donald and Newey (2001) assume that all IVs are exogenous and strong and directly consider the conditional mean squared error of $\hat{\beta}$

$$\forall S \in \mathcal{S} \qquad R_{DN}(S) = \mathbb{E}\left[(\hat{\beta}_S - \beta)^2 | (z_i)_{i=1}^n\right]$$

which the authors approximate using an a priori consistent estimator of β. More precisely,the authors use Nagar (1959) expansions to approximate the bias of different IV estimators and make use of an a priori consistent estimator of β. These expansions are known to be unstable even in the best case scenario and do not hold if the IVs being considered are weak, very weak or endogenous, see Chaudhuri and Zivot (2011). On the other hand Andrews (1999) assumes that all IVs are strong and coins different criteria based on Sargan-Hansen J statistics to pick the largest set of exogenous IVs. Hence the risk the author actually estimate is

$$\forall S \in \mathcal{S} \qquad R_A(S) = \mathbb{E}\left[(y_i - x_i\hat{\beta}_S)z_{iS}'\right] Var\left[z_{iS}(y_i - x_i\hat{\beta}_S)\right]^{-1} \mathbb{E}\left[z_{iS}(y_i - x_i\hat{\beta}_S)\right]$$

up to some normalization to account for $s = |S|_0$. Note that $R_A(S)$ is a normalized in-sample version of $R_{EXO}(S)$.

Risk estimators The risks R_{EXO}, R_{PMSE} and R_{MSE} are unknown and need to be estimated. I consider their cross-validation average estimators[3] denoted as $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$, and $\hat{R}_{MSE}$. To obtain them in practice the following steps can be followed:

Cross-validation average risk estimator

1. Split the original sample into a validation sample of size n_b and a training sample of size $n - n_b$
2. Compute $\hat{\beta}_S$ using the data from the training sample only

[3]In practice other methods such as out-of-bag bootstrap validation or k-fold cross-validation can be used, moreover instead of the average risk the median risk or most voted risk can also be used.

3. Use the validation sample to estimate R_{EXO}, R_{PMSE} and R_{MSE} but plug-in the estimator $\hat{\beta}_S$ created using the training sample

4. Repeat the process B times and average

To be more specific let B be the number of times the original sample is split, n_b be the validation sets sample size, and $n - n_b$ be the training sets sample size. Then for any $b = 1, \ldots, B$ let $(w_i)_{i \in I_b}$ be the validation sample for split b of size n_b and let $(w_i)_{i \in \bar{I}_b}$ be the training sample for split b of size $n - n_b$. Finally, let $\hat{\beta}_{S,b}$ be the 2SLS estimator associated to split b which uses the training sample $\bar{I}_b$ only. Formally for any $S \in \mathcal{S}$ the risk estimators are

$$\hat{R}_{EXO}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_b(s+1)\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - x_i\hat{\beta}_{S,b})z_{iS}'\right) \hat{\Sigma}_S^{-1} \frac{1}{n_b} \sum_{i \in I_b} \left(z_{iS}(y_i - x_i\hat{\beta}_{S,b})\right)$$

$$\hat{R}_{PMSE}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_b\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - z_{iS}'\hat{\pi}_S\hat{\beta}_{S,b})^2\right)$$

$$\hat{R}_{MSE}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_b\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - x_i\hat{\beta}_{S,b})^2\right)$$

where $\hat{\sigma}_b^2$ is a normalization which controls for differences in variations across splits. This normalization is useful in finite sample, for instance $\hat{\sigma}_b^2 = \frac{1}{n_b}\sum_{i \in I_b}^{n_b}(y_i - \overline{y})^2$ or $\hat{\sigma}_b^2 = \frac{1}{n_b}\sum_{i \in I_b}^{n_b}(x_i - \overline{x})^2$, in large samples one can set $\hat{\sigma}_b^2 = 1$. Other types of normalizations are possible for instance adding a degenerate bonus term in s as in Andrews (1999) would allow to select larger subsets of IVs. The practical choice of B, n_b and $n - n_b$ is up to the researcher. A standard choice in machine learning is forty splits with a third of the data used for validation and two thirds of the data used for training, ie $B = 20$, $n_b = \frac{n}{3}$, and $n - n_b = \frac{2n}{3}$. To establish asymptotic results a requirement is that B, n_b, and $n - n_b$ increase with n.

Selection procedure For a certain risk $k \in \{EXO; PMSE; MSE\}$ the selected subset of IVs is simply the minimizer of the risk estimator

$$\hat{S}_{\hat{R}_k} = \underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S)$$

In that sense if K_z is large it becomes very time-consuming to compute $\hat{R}_k(S)$ for all $2^{K_z} - 1$ subsets in $\mathcal{S}$. However, even if $K_z \geqslant 10$ it is still possible to simplify the problem by minimizing the risks over only part of $\mathcal{S}$ or if there are groups of uncorrelated IVs by minimizing the risks in each group as in Windmeijer, Liang, Hartwig, and Bowden (2021).

3.2 Ideal properties of risk estimators

Before deriving the theoretical performances of the selection procedures, I characterize their ideal properties.

Efficiency A selection method $\hat{S}_{\hat{R}_k}$ is deemed efficient if its minimum converges to the minimum of the risk it is trying to estimate

$$\frac{\underset{S\in\mathcal{S}}{min}\ \hat{R}_k(S)}{\underset{S\in\mathcal{S}}{min}\ R_k(S)} \xrightarrow{\mathbb{P}} 1 \tag{3.7}$$

A procedure being efficient does not directly imply that the selection procedure will select a "good" model however. This is especially the case in linear IV models.

Consistency A consistent model selection procedure is a procedure which selects the true model with probability 1 at the limit. But defining a true or good model in the linear IV context is difficult. A candidate set of good models of interest may be set of all IV subsets which allow to identify β

$$\mathcal{S}_{id} = \{S \in \mathcal{S} : \alpha = 0, \pi_S \neq 0\}$$

If $\mathcal{S}_{id}$ is non-empty then there exists at least one valid subset of IVs and therefore β is identified. But identification of β does not guarantee its consistent estimation, in fact a local lack of identification does not prevent consistent estimation and.

Thus, IV subsets of much more interest are the subsets for which $\hat{\beta}_S$ is a consistent estimator of β and the subsets for which the t-statistic $t_S = \frac{\hat{\beta}_S - \beta}{Var(\hat{\beta}_S)}$ is asymptotically standard normal. To characterize these sets, I let a_S and b_S represent respectively the strength of IV z_{iS}, ie $\pi_S \propto n^{-a_S}$, and the level of endogeneity of IV z_{iS}, ie $\mathbb{E}(z_{iS}u_{iS}) \propto n^{-b_S}$. Allowing π_S and $\mathbb{E}(z_{iS}u_{iS})$ to depend on sample size is a way to model IV weakness and IV local endogeneity. This generalization resembles that of Andrews and Cheng (2012) and is a theoretical way to approximate the behavior of IV estimators and inference procedures under unfavorable conditions in practice. Hence, define the three following categories of IV subsets

$$\mathcal{S}_c = \{S \in \mathcal{S} : b_S - a_S > 0, a_S < 1/2\}$$
$$\mathcal{S}_{an} = \{S \in \mathcal{S} : a_S < 1/2, b_S > 1/2\}$$

It can be shown that $\mathcal{S}_c$ represents all the subsets of IVs such that $plim\ \hat{\beta}_S = \beta$ whereas $\mathcal{S}_{an}$ represents all the IV subsets such that $\hat{\beta}_S$ is consistent and asymptotically normal in the sense

that under the null $H_0 : \beta = 0$ the ration $\hat{\beta}_S$ and an estimator of its standard deviation is asymptotically standard normal and therefore the usual Gaussian confidence intervals are valid. Another category of IV subsets of specific interest is

$$\mathcal{S}_r = \{S \in \mathcal{S} : b_S > 1/2\}$$

which characterizes the subsets which yield valid weak identification robust confidence sets for β via test inversion as in Anderson and Rubin (1949). See appendix D.2 and Proposition 4.3 for formal proofs.

Going forward $\hat{S}_{\hat{R}_k}$ is c-consistent where $c \in \{c; an; r\}$ if

$$\mathbb{P}(\hat{S}_{\hat{R}_k} \in \mathcal{S}_c) \to 1 \tag{3.8}$$

Efficiency and consistency of IV selection via $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ are proven in the next section.

Valid post model selection inference Lastly, let $CI_{\alpha,S}(\beta)$ be a confidence interval with nominal coverage α using subset of IVs S based on either weak identification robust methods, either t-tests Gaussian asymptotics. Then, if there exists some $S \in S_r$ or $S \in S_{an}$ valid inference is possible, ie $lim\ \mathbb{P}(\beta \in CI_{\alpha,S}(\beta)) \geqslant \alpha$. Ideally post-selection inference should also be valid however $\hat{S}_{\hat{R}_k}$ is correlated with the data therefore

$$\mathbb{P}(\beta \in CI_{\alpha,\hat{S}_{\hat{R}_k}}(\beta) | \hat{S}_{\hat{R}_k} = S) \neq \mathbb{P}(\beta \in CI_{\alpha,S}(\beta))$$

Valid inference on β post selection via $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ is not formally proven in this paper. But from the extensive simulation exercise in section 5 this contamination does not seem to be a concern as confidence intervals have nominal coverage.

Still, this issue can be completely bypassed and exact inference can be recovered by using sample-splitting[4] with one sample used for finding $\hat{S}_{\hat{R}_k}$ and the other used for estimation and inference. Other common methods for valid-post-selection inference in econometrics and statistics systematically involve immunization of either the inference procedure or the estimation procedure to the choice of nuisance parameter or to the models. For instance, if the estimator is modified by Neyman orthogonalization of the score of the criterion it is built from, as in Chernozhukov et al. (2015) or Singh and Sun (2021) in the IV context, it will

[4]Note that sample splitting will effectively reduce sample size and therefore could aggravate weak IVs problems but at the same time it could improve the level of exogeneity of the IVs.

require all the IVs to be exogenous. A partial identification approach immune to the choice of IVs and therefore endogeneity itself may be used however the confidence interval will be very large unless strong additional assumptions are made. Thus, these approaches are not very appealing, even more so because the IV estimator is taken as given. In fact ideally the estimator and inference procedure should not be immune to the choice of IV subset otherwise it would be impossible to assert the performances of each IV subset.

On a final note $\mathcal{S}_{an}$ being non-empty implies that there exists at least some IVs which are only locally endogenous, if instead $\mathcal{S}_{an}$ is empty valid post-selection inference may be possible for a pseudo-true-value instead.

4 Theoretical properties

To understand the properties of IV selection methods via minimization of $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ defined in section 3.1, the risks are first decomposed then their theoretical asymptotic properties are derived.

4.1 Risks decomposition

In linear models the mean square error of prediction decomposes into the squared bias and variance of estimators and the same decomposition exercise can be performed with R_{EXO}, R_{PMSE} and R_{MSE}, formal proofs are in appendix D.3. For any $S \in \mathcal{S}$ the risks can be rewritten as

$$
\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right) \\
R_{PMSE}(S) &= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right) \\
R_{MSE}(S) &= \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}
$$

Clearly the risks do not decompose into squared bias and variance. Instead, they mainly depend on some average distance between $z_{\overline{E}}^{*'}\alpha$ and $z_S^{*'}\pi_S(\hat{\beta}_S - \beta)$ and it is possible to go further.

Strong and endogenous IVs Assume that IV subset S is strong but endogenous, ie assume that $\pi_S \neq 0$ and is fixed and that $\mathbb{E}(z_{iS}u_{iS}) = \mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha \neq 0$ and is fixed. Then it can be

shown that

$$\hat{\beta} - \beta = \frac{\pi_S' z_S' z_E \alpha}{\pi_S' z_S' z_S \pi_S} + o_P(1) = \frac{\pi_S' \mathbb{E}(z_{iS} z_{i\overline{E}}')\alpha}{\pi_S' \mathbb{E}(z_{iS} z_{iS}')\pi_S} + o_P(1)$$

where $o_P(1)$ is the small o in probability notation[5]. As a consequence, the risks can be rewritten as

$$\begin{aligned}
R_{EXO}(S) &= \alpha' \mathbb{E}(z_{\overline{E}}^* z_S^{*'}) M_1 \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha + o_P(1) \\
R_{PMSE}(S) &= \mathbb{E}\left((u^* - v^*\beta)^2\right) + \alpha' M_2 \alpha + o_P(1) \\
R_{MSE}(S) &= \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \alpha' M_2 \alpha + o_P(1)
\end{aligned}$$

where M_1 and M_2 are positive semi-definite matrices defined as

$$M_1 = \Sigma_S^{-1} - \pi_S(\pi_S' \Sigma_S \pi_S)^{-1} \pi_S', \qquad M_2 = \Sigma_E - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi_S' \Sigma_S \pi_S)^{-1} \pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})$$

From this decomposition it is clear that the risks are mainly quadratic functions of $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$ because $\hat{\beta}_S - \beta$ is also a function of $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$. Thus, the larger the amount of endogeneity or equivalently the larger $\mathbb{E}(z_S^* u_S^*) = \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$ is, the larger are the risks. Note the first term in R_{MSE} also depends on $\hat{\beta}_S - \beta$ thus this risk may miss-classify some IV subsets.

Exogenous IVs This time assume that the right IVs were chosen $S = E$, or that all the IVs are exogenous $\alpha = 0$, then the risks reduce to quadratic terms of $\hat{\beta}_S - \beta$

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(||\Sigma_S^{1/2} \pi_S(\hat{\beta}_S - \beta)||^2\right) \\
R_{PMSE}(S) &= \mathbb{E}\left((u^* - v^*\beta)^2\right) + \mathbb{E}\left(||z_S^{*'} \pi_S(\hat{\beta}_S - \beta)||^2\right) \\
R_{MSE}(S) &= \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_S^{*'} \pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

Thus, if the subset S is strong then $\hat{\beta}_S$ is a consistent estimator of β and the three risks further reduce asymptotically, whereas if the subset S is weak then $\hat{\beta}_S$ does not converge to β and the three risks are large.

To summarize, when ranking IV subsets the risks are not weighting subsets to balance bias and variance, instead they are larger if IVs are endogenous and slightly larger if the IVs

[5]Formally if random sequence $X_n = o_P(1)$ then $\forall \varepsilon > 0 \ \mathbb{P}(|X_n| > \varepsilon) \to 0$. If random sequence $X_n = O_P(1)$ then $\forall \varepsilon > 0 \ \exists M > 0, \exists N > 0 : \forall n > N \ \mathbb{P}(|X_n| > M) < \varepsilon$.

are weak. In that sense endogeneity is first order whereas weakness is second order, this is a desirable property because valid inference on β can still be performed regardless of the level of strength as long as the IVs being picked are exogenous.

4.2 Asymptotic results

In order to establish the efficiency of the selection procedures three sets of assumptions are made. Assumption A characterizes the model and the data generating process (DGP). Assumption B places restrictions on the k-class of IV subset estimators. Assumption C characterizes the risk estimator to make sure it convergences towards the risk.

Assumption A
(i) The sample $(y_i, x_i, z_i)_{i=1}^n$ *is iid such that (2.1) and (2.2) hold at* β
(ii) z_i, x_i *and* y_i *possess finite moments of order* 4, z_i *is not perfectly colinear*
(iii) Without loss of generality for any $S \in \mathcal{S}$, $\pi_S \equiv n^{-a_S}\kappa_S$ *for some fixed* $\kappa_S \in \mathbb{R}_*^s$ *and some* $a_S \in \mathbb{R}^+ \cup \{+\infty\}$
(iv) Without loss of generality for any $S \in \mathcal{S}$, $\mathbb{E}(z_{iS}u_{iS}) = \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = n^{-b_S}\delta_S$ *for some fixed* $\delta_S \in \mathbb{R}_*^s$ *and some* $b_S \in \mathbb{R}^+ \cup \{+\infty\}$

Assumption A(i) determines the model whereas A(ii) is a common moments condition. Alternatively, conditional heteroskedasticity could be assumed. A(iii) and A(iv) formally allow the level of weakness and the level of endogeneity of any subset of IVs to vary with n, thus the asymptotic behavior of the IV estimator and risk estimators is characterized by the values of (a_S, b_S). For instance, when $a_S \geqslant 1/2$, IV subset S is weak and therefore the estimator $\hat{\beta}_S$ is random at the limit. See lemma in appendix C for details.

Assumption B
For any $S \in \mathcal{S}$ *there exists some* $e > 0$ *such that*

- *If* $a_S \geqslant 1/2$
$$\mathbb{P}(x' P_{z_S} x < e) = 0$$
- *If* $a_S < 1/2$
$$\mathbb{P}(n^{2a_S - 1} x' P_{z_S} x < e) = 0$$

Assumption B is a condition which ensures the existence of moments of the 2SLS estimator (more precisely, they ensure the uniform integrability of the risk estimators). B is almost always satisfied in practice, for instance if x_i is continuous, or if it is discrete but doesn't have zero in its support. An unnatural counterexample would be the case where x is binary with a very large probability to be equal to zero.

Assumption C

For $k \in \{EXO; PMSE; MSE\}$ let $\hat{R}_{k,b}(S)$ be the risk estimator computed for split b. Then n_b and B are such that

(i) $n_b \underset{n\to+\infty}{\to} +\infty$, $n - n_b \underset{n\to+\infty}{\to} +\infty$, and $B \underset{n\to+\infty}{\to} +\infty$

(ii) There exists some $c \in (0;1)$ such that for $k \in \{EXO; PMSE; MSE\}$, for any $S \in \mathcal{S}$, for any $b = 1, \dots, B$

$$\sum_{b'=1}^{B} Cov(\hat{R}_{k,b}(S), \hat{R}_{k,b'}(S)) \leqslant \sum_{n_t=0}^{n_b} Var(\hat{R}_{k,b}(S)) c^{n_b - n_t}$$

Assumption C characterizes the sampling process used to obtain the risk estimators. This assumption is specific to the linear IV context with potentially endogenous and potentially weak IVs. As such, from C(i) the training sample size and validation sample size need to increase with n, this implies that leave-one-out cross-validation cannot be used for risk estimation. This is necessary in order to estimate the out-of-sample correlation between the IVs and the error. C(ii) is a sufficient condition to ensure convergence of the average risk and can be ignored when all IVs are strong. Intuitively, it forces the correlation between the estimated risk across splits to be proportional to the number of common observations across splits. Consequently, even when 2SLS is random at the limit because the IVs are weak, the estimators across splits are not too correlated. A simple way to satisfy C(ii) which is quite common in the machine learning literature is to randomly split the data into B folds of size n/B and split those B folds again into training and validation samples. This way, the risk estimators across folds are effectively independent. Note that C(ii) is always satisfied in practice from simulation evidence (the constant c can be close to 1).

Theorem 4.1 states that under the above assumptions IV subset selection via cross-validation is efficient, its proof is in appendix C.

Theorem 4.1

Under assumptions A, B, and C, for $k \in \{EXO, PMSE, MSE\}$

$$\frac{\min\limits_{S \in \mathcal{S}} \hat{R}_k(S)}{\min\limits_{S \in \mathcal{S}} R_k(S)} \xrightarrow{\mathbb{P}} 1$$

An almost sure version of this result cannot be obtained unless one assumes that all IVs are strong ($a_S < 1/2$). This is due to the fact the 2SLS estimator is random at the limit when IVs are weak ($a_S \geqslant 1/2$). As mentioned, efficiency in this context does not guarantee the selection of good IV subsets.

Consistency can be established with another assumption.

Assumption D

(i) If $k = MSE$ *then at least one of the following conditions must hold for any* S

- $sign(\rho) \neq sign(\delta_S' \kappa_S)$
- $\delta_S' \kappa_S = 0$

(ii) Let K_w *denote the dimension of the largest subset* $S \in \mathcal{S}$ *such that* $a_S \geqslant 1/2$ *then at least one of the following conditions hold*

- $K_w \leqslant 2$
- $\frac{\sigma_u^2 \sigma_v^2}{\rho^2} > max\{K_w - 1, \frac{K_w}{2}, \frac{K_w(K_w-1)}{2}\}$

Assumption D(i) ensures that R_{MSE} will correctly rank subsets with varying levels of endogeneity. As mentioned in section 4.1, R_{MSE} decomposes into a first term $\mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right)$ and a term which is quadratic in $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$. Under specific conditions, the dependence of the first term on $\hat{\beta}_S - \beta$ may lead R_{MSE} to miss-classify some IV subsets, typically when the sign ρ of the OLS endogenous bias is the same as the endogenous bias $\delta_S' \kappa_S$ of 2SLS using IV subset S. D(ii) ensures that when 2SLS is used the three risks will correctly rank strong subsets below weak subsets. Indeed, 2SLS is close to OLS in terms of behavior when IVs are weak and numerous, in fact it can be shown that on average 2SLS behaves like OLS in such conditions therefore on average it can be very efficient and therefore misclassified. Consequently, D(ii) is unnecessary when using other k-class estimators due to their lower bias.

Theorem 4.2 establishes consistency under the above assumptions, its proof is in appendix C.

Theorem 4.2
Under assumptions A, B, C and D

- *If* $\mathcal{S}_c \neq \emptyset$ *and assumption D hold then for* $k \in \{EXO, PMSE\}$

$$\mathbb{P}(\underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_c) \to 1$$

- *If* $\mathcal{S}_{an} \neq \emptyset$ *then for* $k \in \{EXO, PMSE, MSE\}$

$$\mathbb{P}(\underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_{an}) \to 1$$

- *If* $\mathcal{S}_r \neq \emptyset$ *then for* $k \in \{EXO, PMSE, MSE\}$

$$\mathbb{P}(\underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_r) \to 1$$

Intuitively, if there exists some IV subset such that the estimator is consistent or asymptotically normal then one such subset will be picked with probability one at the limit. Similarly, if there exists some IV subset such that valid weak identification robust inference can be performed then one of such subset will be picked with probability at the limit. This result has some caveats, however. If the mean square error of prediction R_{MSE} is used to pick the IV subset, then it may not detect endogeneity in certain situations. If the 2SLS estimator is used, it may lead the risks to not detect weak IVs if they are too numerous.

It is very difficult to establish more precise results regarding the exact identity of the IV subsets being picked. Such results could be established by either assuming normality of the data and that IVs are non-random, which are very unrealistic assumptions, or assuming that all IVs are strong and use Nagar expansions as in Nagar (1959), which are known to be unreliable. In the next section, I assess the empirical performances of the IV subset selection procedures.

5 Simulations

In this section, I perform an extensive simulation exercise in order to assess the behavior of different IV selection procedures in different settings: The general case where IVs can be endogenous and / or weak; The case where IVs are strong but can be endogenous IVs are exogenous but some may be weak; The case where IVs are exogenous but can be weak.

Performance Measures I evaluate the performances of the selection methods through the performances of the 2SLS estimator post selection and a weak identification robust confidence interval based on the conditional likelihood ratio test with normal approximation of Mikusheva (2010) post selection. Thus, I consider the following metrics over 20,000 simulations: The interquartile range of the estimators across simulations; The median absolute bias; The median squared bias; The empirical coverage of β using normal asymptotics for inference with nominal coverage 95%; The median length of the confidence interval (CI) using normal asymptotics; The average number of IVs being picked; The empirical coverage of β using the weak identification robust confidence interval (RCI); The median length of the the weak identification RCI; And the percentage of times the weak identification robust confidence interval exists and is finite.

Selection methods The risk estimators from section 3.1 are computed using the cross-validation risk estimators with $B = 40$ resamples and $n_b = n/2$. Furthermore, to make comparisons, I compute the same post selection metrics using the following selection methods: The "mean square error" criterion of Donald and Newey (2001) using the jacknife and the 2SLS using all the IVs as a first stage; The GMM-BIC procedure of Andrews (1999); The post-lasso of Kang et al. (2016) with the penalty obtained by cross-validation; The post-adaptive-lasso of Windmeijer et al. (2018) with the penalty obtained by cross-validation and using the median estimator as a first stage; The oracle which only uses strong and exogenous IVs. To extend on the introduction, Donald and Newey (2001) should fail in case some IVs are endogenous, Andrews (1999) should fail in case some IVs are weak, Kang et al. (2016) should fail unless the majority of IVs are exogenous and all IVs are strong, and Windmeijer et al. (2018) should fail unless the "largest group" of IVs is exogenous and all IVs are strong. Note that the Lasso and post-lasso are much better suited to a setting with a larger number of IVs unlike the methods developed in this paper.

Data generating process The data generating process I consider through the simulations is of the following form

$$\begin{aligned} y_i &= 2x_i + \alpha_2 z_{i2} + \alpha_4 z_{i4} + \alpha_6 z_{i6} + u_i \\ x_i &= 0.5\left(z_{i1} + z_{i2} + \frac{c_1}{\sqrt{n}} z_{i3} + \frac{c_1}{\sqrt{n}} z_{i4} + \frac{c_2}{n} z_{i5} + \frac{c_2}{n} z_{i6}\right) + v_i \end{aligned} \tag{5.9}$$

where $(u_i, v_i, z_i')_{i=1}^n$ is iid, normally distributed with mean 0, individual variance 1, correlation between IVs is equal to 0.1, and correlation between u_i and v_i is equal to 0.5. Thus, there are

six IVs, with three which potentially do not satisfy the exclusion restriction. The parameters which determine the different settings are $(c_1, c_2) \in \mathbb{R}^2$ and $(\alpha_2, \alpha_4, \alpha_6) \in \mathbb{R}^3$. I also allow the sample size n to be equal to either 400 or 4000.

Note that using this specification some IV subsets end up at the exact cutoff levels in terms of endogeneity and strength which determine the asymptotics of the risks, and some IVs have exactly the same level of strength, therefore it constitutes a worst-case scenario for the selection procedures. Moreover, this specification also resembles the simulation designs in Belloni et al. (2012). In addition, this specification is of specific interest because it allows to control for the range of IVs strength and for the level of bias in terms of pseudo-true-value for all the possible sets of IVs. In each setting I mention the bias of the OLS estimator, the range of the concentration parameters divided by the number of IVs and the range of the (pseudo-true-value) bias of the 2SLS estimator over all the possible subsets S in $\mathcal{S}$. The bias of the OLS estimator can be written as

$$Bias(OLS) = \left| \frac{\rho + \pi' \mathbb{E}(z_i z'_{i\overline{E}})\alpha}{\mathbb{E}(x_i^2)} \right|$$

The pseudo-true-value bias due to using subset S as IVs is

$$Bias(Pseudo_S) = \left| \frac{\pi'_S \mathbb{E}(z_{iS} z'_{i\overline{E}})\alpha}{\pi'_S \mathbb{E}(z_{iS} z'_{iS})\pi_S} \right|$$

The concentration parameter of set S is a measure of strength of the IVs S which is defined as

$$\mu_S^2 = \frac{n\pi'_S \mathbb{E}(z_{iS} z'_{iS})\pi_S}{s\sigma_v^2}$$

It is typically considered low and IVs are typically considered jointly weak when it is inferior to 20, see Stock et al. (2002).

5.1 General case

The general setting I consider is such that $c_1 = 1$, $c_2 = 1$, and $\alpha_2 = \alpha_4 = \alpha_6 = 1$. Thus, model 5.9 can be rewritten as

$$y_i = 2x_i + z_{i2} + z_{i4} + z_{i6} + u_i, \qquad x_i = 0.5\left(z_{i1} + z_{i2} + \frac{1}{\sqrt{n}} z_{i3} + \frac{1}{\sqrt{n}} z_{i4} + \frac{1}{n} z_{i5} + \frac{1}{n} z_{i6}\right) + v_i$$

In addition, $Bias(OLS) = 0.83$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0; 800]$ and $\mu_S^2 \in [0; 100]$ for $n = 400$ whereas $Bias(Pseudo_S) \in [0; 8000]$ and $\mu_S^2 \in [0; 1000]$ for $n = 4000$. Consequently,

there are some very weak and endogenous IV subsets and some strong and exogenous IV subsets. The IV subsets which will yield an asymptotically normal estimator are therefore composed of z_{i1} and possibly of z_{i3} and z_{i5}

$$\mathcal{S}_{an} = \{z_{i1}; \{z_{i1}; z_{i3}\}; \{z_{i1}; z_{i5}\}; \{z_{i1}; z_{i3}; z_{i5}\}\}$$

The oracle only uses z_{i1} for instrumentation.

Simulations results for 2SLS in this general setting are in table 1 of appendix A.2. The criterion of Donald and Newey (2001) balances bias and variance assuming exogeneity of all IVs and therefore picks a subset of strong IVs regardless of their level of endogeneity. Consequently, the estimator has high bias, very low coverage, but is efficient. Andrews (1999) fails to pick any decent subset of IVs because it requires all of them to be strong, and for there to be no corner solutions. Its estimates post-selection are extremely biased and are very different across simulations, this indicates that weak IVs were picked. Lastly, the post-lasso of Kang et al. (2016) and the post-adaptive-lasso of Windmeijer et al. (2018) have high dispersion, which means that for each simulation very different IV subsets are picked. As a consequence, the performance of the post-lasso and post-adaptive-lasso are not great, there is bias and coverage is low. This is due to the fact that IVs are allowed to be weak and that the exogeneity conditions required for these procedures to work do not hold. The three risks R_{EXO}, R_{PMSE}, and R_{MSE} are performing as well as the oracle in large sample in terms of bias, coverage and length of Gaussian confidence interval, and dispersion. Note that R_{PMSE} picks more IVs than the other two methods. In addition, in smaller samples, some coverage seems to be lost especially when R_{EXO} is used for selection.

Simulations diagnostics for the RCI using the CLR test are in the first three columns of table 5 in appendix A.2. As with the performances of the post selection 2SLS estimator, the RCI using R_{EXO}, R_{PMSE}, and R_{MSE} for selection performs at oracle level in terms coverage, interval lengths, and percentage of finite confidence interval. In small sample this is less the case, using R_{MSE} the RCI undercovers and using R_{EXO} the RCI can be of infinite length. Other selection methods perform worse in terms of coverage and interval length.

5.2 Strong IV case

I consider two strong IV specifications. In the first strong IV setting I let $c_1 = \sqrt{n}$, $c_2 = n$, $\alpha_2 = \alpha_4 = \alpha_6 = 1$ and model 5.9 can be rewritten as

$$y_i = 2x_i + z_{i2} + z_{i4} + z_{i6} + u_i, \qquad x_i = 0.5 \sum_{k=1}^{6} z_{ik} + v_i$$

Thus, $Bias(OLS) = 0.84$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0;2]$ and for $n = 400$ $\mu_S^2 \in [100;150]$ whereas for $n = 4000$ $\mu_S^2 \in [1000;1500]$. Hence, all IV subsets are strong, some are very endogenous however, and this time $\mathcal{S}_{an}$ is composed of any combination of z_{i1}, z_{i3} and z_{i5}. The oracle uses (z_{i1}, z_{i3}, z_{i5}).

Simulations results for 2SLS in this strong IV setting are in table 2 of appendix A.2. Again Donald and Newey (2001) selects strong IVs thus it picks all six IVs because all of them are strong leading to an efficient but very biased estimator and to a confidence interval with coverage 0. Andrews (1999) J statistic criterion fails to pick the exogenous IVs, thus the 2SLS estimator post-selection is still highly biased, and coverage of the Gaussian confidence interval is low. This is most likely due to the fact that in this setting the J test statistic is low even when IVs are endogenous, indeed in many settings the J test has low power under the alternative, see Kiviet and Kripfganz (2021) for a recent review and for formal conditions under which the J test has no power. As before the post-lasso and post-adaptive Lasso perform badly, this is due to the fact that the IVs do not satisfy the right exogeneity conditions. Thus, only R_{EXO}, R_{PMSE} and R_{MSE} are selecting strong IV subsets and have performances comparable to the oracle with a large sample, and slightly lower performances in small sample.

The second strong IV setting which I call favorable setting is one where a strict majority of IVs is exogenous and where the levels of endogeneity of the endogenous IVs are different from each other and large. In such setting the Sargan Hansen J statistic will be large if endogenous subsets are picked, the median IV estimator which is used to set up the weights in the adaptive Lasso procedure is consistent, and there can be no confusion between weak and exogenous IV subsets and weak and endogenous IV subsets in the Lasso procedure. Let $c_1 = \sqrt{n}$, $c_2 = n$, $\alpha_2 = 0$, $\alpha_4 = 1$, and $\alpha_6 = 3$ and 5.9 can be rewritten as

$$y_i = 2x_i + z_{i4} + 4z_{i6} + u_i, \qquad x_i = 0.5 \sum_{k=1}^{6} z_{ik} + v_i$$

Then $Bias(OLS) = 1.07$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0;6]$, for $n = 400$ $\mu_S^2 \in [100;150]$ and for $n = 4000$ $\mu_S^2 \in [1000;1500]$. The oracle uses $(z_{i1}, z_{i2}, z_{i3}, z_{i5})$.

Simulations results for 2SLS in this strong and favorable setting are in table 3 of appendix A.2. All the selection procedures are performing well except Donald and Newey (2001) which does not recognize endogenous IVs. Note that the 2SLS estimators obtained after selection via Lasso and adaptive Lasso are slightly more dispersed across simulations, have higher bias and their Gaussian confidence intervals have lower coverage compared to estimators after

selection using Andrews (1999), R_{EXO}, R_{PMSE} or R_{MSE}. This is not surprising, Lasso and adaptive Lasso can be useful when the total number of IVs is large and when sample size is large. When the total number of IVs is low there is few reason to use Lasso, just like there is no reason to use Lasso for control variable selection in linear models when there aren't many control variables in the first place.

In these two strong IV settings, simulations diagnostics for the RCI using the CLR test are in the fourth to ninth columns of table 5 in appendix A.2. Once again because Donald and Newey (2001) picks endogenous IVs, using it to select IVs yields an RCI with very low coverage in both strong IV settings. Andrews (1999) doesn't pick exogenous IVs in the first setting whereas it does in the second, this is reflected in the performances of its RCI which is at oracle level in the second setting but not in the first. The RCI after Lasso or adaptive-lasso selection perform very badly in the first setting for the same reasons post-lasso and post-adaptive-lasso perform badly, in the setting they perform very well but not at oracle level. Lastly, the RCI using R_{EXO}, R_{PMSE}, and R_{MSE} for selection have oracle level performances in both settings.

5.3 Exogenous IV case

Lastly, I consider an exogenous IVs setting such that $c_1 = c_2 = 1$ and $\alpha_2 = \alpha_4 = \alpha_6 = 0$

$$y_i = 2x_i + u_i, \qquad x_i = 0.5\left(z_{i1} + z_{i2} + \frac{1}{\sqrt{n}}z_{i3} + \frac{1}{\sqrt{n}}z_{i4} + \frac{1}{n}z_{i5} + \frac{1}{n}z_{i6}\right) + v_i$$

Then $Bias(OLS) = 0.32$, whereas $Bias(Pseudo_S) = 0$ for any $S \in \mathcal{S}$, $\mu_S^2 \in [0;100]$ for $n = 400$ and $\mu_S^2 \in [0;1000]$ for $n = 4000$. Therefore, there are very weak sets of IVs and strong sets of IVs and as long as a set of strong IVs is picked the estimator should estimate the true causal parameter of interest. The oracle use (z_{i1}, z_{i2}) for instrumentation.

Simulation results for 2SLS in this exogenous setting are in table 4 of appendix A.2. In this case, there are very little differences in terms of estimator performances between each selection method, it seems all of them are picking strong IVs and some of them also pick weaker IVs. Post-lasso seem to select too few IVs which is why it seems less efficient. On another note, in principle Andrews (1999) picks the largest IV subset which is exogenous, this explains why it picks all six IVs. As for simulations diagnostics for the RCI using the CLR test they are in the tenth to twelfth columns of table 5 in appendix A.2. All methods yield a post selection RCI with oracle performances except, as argued previously, Lasso which select too few IVs and thus has an RCI with a slightly greater length.

In the next section I estimate the effect of pre-trial detention on guilt and I select judge dummies which act as instrumental variables with the methods designed in this paper.

6 Application

Since Kling (2006) a large literature in Economics and Law which utilizes the random assignment of judges to cases and differences in the degree of severity of judges has developed in order to estimate the causal effect of prison on offenders' outcomes. It is now well-established that, controlling for the offender's characteristics, for the case's characteristics, for other time and place variables, judges differ significantly in their propensity to send offenders to prison, including their propensity to send offenders to pre-trial detention. Thus, in practice judge dummies generate supposedly exogenous variation in detention / pre-trial detention which allow to identify and estimate causal effects, most often the JIVE estimator is used instead of 2SLS leading to the famous judge leniency IV or jacknifed judge IV. This identification strategy can fail for multiple reasons, however.

First, judges may not differ significantly in their leniency, which can generate a weak IVs' problem. Second, the identity of the judge assigned to a case (and his level of leniency) is known by the offender and his lawyer before the trial is held, this can lead to voluntary postponement of the trial in order to get a more lenient judge, bribing of the judge, differing defenses during the trial depending on the judges' leniency, plea deals before the actual trial, etc... Furthermore, when evaluating the effect of pre-trial detention on detention, the identity of the judge present during the pre-trial hearing is also known by the judge present for the actual trial. Consequently, the judge present during the trial has the possibility of doubling-down on the signal sent by the judge during the pre-trial hearing if they deem them trustworthy, or on the contrary compensate for the pre-trial judgement if they deem the judge present during the pre-trial hearing untrustworthy. Finally, judges are never completely randomly assigned to a case, at best a judge among a subset of available judges is randomly assigned to a case. As a consequence, the identity of the judge can directly affect the offender outcomes or indirectly through unobserved cofounders such as offender income, level of education, lawyer quality, defense strategy, psychological state.

For the aforementioned reasons, selecting the judges which differ most in their leniency and which satisfy best the exclusion restriction is a priority. In this specific application I use data on 331,971 court cases in Philadelphia and Miami from September 2006 until February 2013 and study the effect of pre-trial detention on the likelihood of being found guilty using

8 judge dummy variables as IVs. Among other control variables the data includes the case characteristics, the offender criminal history and some of their characteristics such as their race, the date and time of the day when the pre-trial hearing is held. Most information about the actual trial are unknown, including the identity of the judge present during the trial. To be more specific, after their arrest an offender is assigned a judge who will preside over a pre-trial hearing, in Philadelphia and Miami pre-trial hearings happen at most a few days after initial arrest. During this hearing, the judge decides whether they offer the offender a plea deal, whether they offer the offender a bail deal, or whether they directly send the offender to prison before their trial. Note that because the data is from the US, failure of the exclusion restriction due to bribing is unlikely. Heterogenous effects are also unlikely because pre-trial detention sends the same signal to the judge which presides over the actual trial. Finally, from table 7 in appendix A.3 the observable covariates seem relatively balanced across judges. See Stevenson (2018) for more details on the data and the set-up.

To be more specific, the oracle model to estimate is

$$guilt_i = predet_i\beta + X_i'\delta + \sum_{j\in E} \alpha_j 1\{judge_i = j\} + u_i$$

$$predet_i = \sum_{j=1}^{8} \pi_j 1\{judge_i = j\} + X_i'\gamma + v_i$$

where $guilt_i$ is a dummy variable which equals 1 if the offender was found guilty and equals 0 otherwise, $predet_i$ is a dummy variable which equals 1 if the offender was kept in prison before his trial and equals 0 otherwise, for $j = 1,\dots,8$ $1\{judge_i = j\}$ is a dummy variable which equals 1 if judge j is the judge who oversees the pre-trial hearing of individual i and equals 0 otherwise, X_i is a vector of control variables which include an intercept, and there exists some j such that $\alpha_j = 0$. Of course a priori it is unknown which judge dummy enters the structural equation, ie $\overline{E}$ is unknown, which is why the judge dummy IVs must be selected.

In table 6 in appendix A.3 are pre-trial judges' descriptive statistics including what percentage of cases they oversaw and what percentage of offenders were sent to prison before their trial unconditionally and conditionally on different control dummy variables. All 8 judges have supervised a high number of cases but some more than others, across all judges offenders are sent to prison before their trial between 39% and 44% of the time. When conditioning on offender characteristics and criminal record such as race, gender or the number of prior offense, the differences in judge propensities to send offenders to prison before the trial are maintained. On the other hand, when conditioning on the case characteristics such

as the time when the offender was arrested or the reason for their arrest, differences in judge propensities for pre-trial detention become very different. This clearly indicates that judges greatly differ in their leniency, that offenders demographics enter linearly in the first stage as fixed effects but that case characteristics do not.

In table 8 in appendix A.3 are the first stage estimates, heteroskedasticity robust standard errors, and relevance p-values, of the judge dummy variables on the endogenous dummy variable pre-trial detention for three different specifications: (1) only includes time and date fixed effects, (2) also includes the case characteristics, and (3) also includes the offender characteristics and their criminal history. First stage F-statistics are also reported, and so are Sargan-Hansen J statistics after using the 2SLS estimator with the full set of IVs. Note that because an intercept is included a judge dummy variable must be excluded to prevent multicolinearity, I excluded judge 2. As a consequence, the judge fixed effects on pre-trial detention are all relative to judge 2. This choice is made mainly because judge 2 oversees the lowest amount of cases which limits potential endogeneity bias, and because judge 2 has the highest probability to send anyone to pre-trial detention which increases the amount of strong IV subsets. A thorough explanation and analysis of the choice of excluded judge and its impact along with a guess on which IV is likely to be endogenous are in appendix B. After excluding judge 2, for any specification the effect of judge 6 is insignificant, ie judge 6 is as harsh as judge 2. On the other hand, judge 1 and judge 5 seem to have the same moderate effect on the likelihood of the offender going to pre-trial detention. Judge 4 and judge 8 are significantly are the most lenient judges. Taken jointly the IVs are not weak but are not very strong either given that the first stage F statistic is close to 40 across all specifications. Consequently, small size and power distortions of tests are to be expected. Regarding the Sargan-Hansen J statistics, they cannot tell us whether or not the IVs are exogenous. As mentioned, the Sargan-Hansen test is known to have poor power properties, see Kiviet and Kripfganz (2021), and indeed specification (1) with the fewest controls is the least likely to satisfy the exclusion restriction and yet it has the lowest J statistic compared to the J statistics in specification (2) and (3) which both reject exogeneity at level 10%.

Next, in table 9 of appendix A.3 I report the OLS estimator, the 2SLS estimator using all the IVs, the 2SLS estimator after selection via Donald and Newey (2001) (DN), via Andrews (1999) (AN), via Kang et al. (2016) (Post-lasso), via Windmeijer et al. (2018) (Post-adalasso), and my methods along with their heteroskedasticity-robust standard errors. I also report the set of judges selected for instrumentation IVs by each method. Thus, the judges which are

not included as IVs were used as control variables except for judge 2 which was excluded. In addition, in table 10 are the first stage F statistics and the J statistics computed post selection for each method.

From table 9 for the three specifications the OLS estimate of the effect of pre-trial detention on guilt is close to zero. This is in accordance with the literature, it is believed that OLS underestimates the effect of pre-trial detention on the likelihood of being found guilty due to omitted variables. Indeed, variables such as the offender level of education or the offender income are unobserved and negatively correlated with pre-trial detention. On the other hand, the 2SLS estimates using all the judge dummies as IVs lie between 15% and 18.5% in the three specifications which is slightly lower than what is expected in the literature, see Kling (2006) or Dobbie, Goldin, and Yang (2018). Regarding the selection procedures, quite interestingly the AN and DN use all the judges as IVs in the three specifications. A possible explanation is that all IVs are exogenous and strong, but based on the J test statistics in table 10 the full set of IVs is unlikely to be exogenous. Thus, the most plausible explanation is that the conditions for these two procedures to work are not met. Next, note that the Lasso uses judge 4 as the sole IV and the 2SLS estimator is approximately equal to 0.25 in all specifications. Judge 4 is both the most lenient judge and is the judge with the highest number of cases, so this IV could be endogenous. This cannot be confirmed by the implied J statistic in table 10 because they are equal to zero by construction. Conceptually, the Lasso will work if a majority of the IVs are exogenous. Thus, it is not clear why it picks the same judge dummy in (1) and (3) even tough the judge dummies in (1) are most likely endogenous. The selection via adaptive Lasso is difficult to interpret, it uses judge 4 and 6 in specification (1), all judges except judge 1 as IVs in specification (2), and judges 6 in specification (3). The post adaptive Lasso 2SLS estimates vary between 0.20 and 0.52 and are significant except in setting (3), and the J statistics are small. In addition, the selection procedure is quite sensitive to the choice to the penalty choice, different non-nested IV sets can be picked depending on the penalty value. A plausible explanation is that there is no "largest group" of exogenous IVs. In addition, the adaptive Lasso can fail when some IVs are weak, which is the case for judge 6 which is selected by the adaptive Lasso in (2) and (3).

Regarding the procedures developed in this paper R_{EXO} selects judges 6 and 7 in specifications (1) and (2), and judges 4 and 8 in specification (3). This choice in specifications (1) and (2) seem to be linked to the fact that there are not enough control variables and therefore that the true effect of pre-trial detention on guilt cannot be properly captured. This would also explain why the J statistics are so low, the procedure may have focused on picking the most

exogenous IV subsets possible. In (3) it selects a completely different set of judges, the effect is significant and close to 0.26, and the J statistic is very small. R_{PMSE} selects judge 4 in (1) and (3) and selects judges 4 and 8 in specification (2). Similar IV subsets are selected across specifications and are very close to that of Lasso. The estimates post-selection with R_{PMSE} are close to 0.25 and are significant in all specifications. Finally, R_{MSE} selects judges 1, 3, 7 and 8 in specification (1) and (2), and judges 1 and 6 in specification (3). The corresponding estimators are not statistically significant, indeed the corresponding first F statistics are not large (< 30). In addition, the estimator is negative in specification (3). This is most likely due to the fact that the sign conditions for R_{MSE} to rank properly IV subsets are not satisfied. Indeed, the OLS estimator is smaller than the 2SLS estimator thus the OLS bias is negative, ie $\rho < 0$. Additionally, as mentioned some judges which supervise pre-trials may be positive signal for the judge during the actual trial. Hence, judge dummies which enter the structural equation may have a positive effect on the likelihood of being found guilty, ie $\mathbb{E}(z_{iS}z_{i\bar{E}})\alpha > 0$. This would violate assumption D(i).

The estimates post-selection with R_{EXO} and R_{PMSE} in specification (3) appear quite trustworthy compared to other procedures, especially in light of the simulation exercise. This means that the effect of pre-trial detention on the likelihood of being found guilty is actually equal to 25%. This is more in line with the literature compared to a 2SLS estimator equal to 18% found using all the judge dummies.

7 Concluding Remarks

In this paper I formally define and study losses, risks, and risk estimators for the selection of strong and exogenous subsets of IVs in the linear IV model for the 2SLS estimator. To do so, I utilize the losses implicitly minimized during IV estimation and obtain three risk: one based on the exogeneity condition, one based on the mean square error of prediction after projecting the endogenous variable on the IVs, and the mean square error of prediction. These risks do not balance squared bias and variance, instead they trade in priority endogenous IVs for exogenous ones. I show that choosing the IV subsets which minimize these risks is a consistent procedure to obtain an estimator which converges towards the true structural parameter of interest and is asymptotically normal. This implies that in practice, applied researchers can use this IV selection method to easily strengthen the credibility of their results. If they have multiple IVs at their disposal they have theoretical guarantees that they will select a strong and exogenous subset. These results are corroborated by the simulation exercise and the

application.

From a broader perspective, this paper resembles earlier works on the selection of regressors via risk minimization when the number of regressors is finite. It is no surprise then that, given a fixed and small total number of IVs, the risks developed in this paper have better chances to select exogenous IV subsets compared to the Lasso and adaptive Lasso which are best suited to a setting with many IVs. A next natural step in the investigation of risks in IV models is understanding when the procedures developed in this paper might fail. Introducing heteroskedasticity, non-linearity, or heterogeneity in the structural equation may affect the selection methods in which case they would require correction. In fact it may be possible to derive new predictions losses and risks for non-linear IV models. These risks could then be used to select the tuning parameters of regularized two-step IV estimators. For instance, the penalty terms of the Lasso and adaptive Lasso could be chosen using risks similar to the ones coined in this paper.

Bibliography

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. K. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–564.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K., V. Marmer, and Z. Yu (2019): "On optimal inference in the linear IV model," *Quantitative Economics*, 10, 457–485.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quaterly Journal of Economics*, 132, 1553–1992.

Antoine, B. and P. Lavergne (2022): "Identification-robust nonparametric inference in a linear IV model," *Journal of Econometrics*.

Apfel, N. and F. Windmeijer (2022): "The Falsification Adaptive Set in Linear Models with Instrumental Variables that Violate the Exogeneity or Exclusion Restriction," *arXiv:2212.04814*.

Arlot, S. and A. Celisse (2010): "A survey of cross-validation procedures for model selection," *Statistics Surveys*, 4, 40–79.

Bai, J. and S. Ng (2010): "Instrument Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26, 1577–1606.

Bates, S., T. Hastie, and R. Tibshirani (2021): "Cross-validation: what does it estimate and how well does it do it?" *arXiv:2104.00673*.

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

Caner, M. (2009): "Testing, Estimation in GMM and CUE with Nearly-Weak Identification," *Econometric Reviews*, 29, 330–363.

Carrasco, M. (2012): "A regularization approach to the many instruments problem," *Journal of Econometrics*, 170, 383–398.

Carrasco, M. and M. Doukali (2021): "Testing overidentifying restrictions with many instruments and heteroscedasticity using regularised jackknife IV," *The Econometrics Journal*, 25, 71–97.

Carrasco, M. and G. Tchuente (2016): "Efficient Estimation with Many Weak Instruments Using Regularization Techniques," *Econometric Reviews*, 35, 1609–1637.

Chaudhuri, S. and E. Zivot (2011): "A new method of projection-based inference in GMM with weakly identified nuisance parameters," *Journal of Econometrics*, 164, 239–251.

Chen, J., D. L. Chen, and G. Lewis (2021): "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," *arXiv:2011.06158*.

Chernozhukov, V., C. Hansen, and M. Spindler (2015): "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, 105, 486–490.

Dobbie, W., J. Goldin, and C. Yang (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–40.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

Gautier, E. and C. Rose (2021): "High-dimensional instrumental variables regression and confidence sets," *Working paper*.

Graddy, K. (2006): "Markets: The Fulton Fish Market," *Journal of Economic Perspectives*, 20, 207–220.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Hahn, J. and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, 163–189.

——— (2003): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review*, 93, 118–125.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations," *The Econometrics Journal*, 7, 272–306.

Hall, A. R. and F. P. M. Peixe (2003): "A Consistent Method for the Selection of Relevant Instruments," *Econometric Reviews*, 22, 269–287.

K. Newey, W. (1985): "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229–256.

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016): "Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization," *Journal of the American Statistical Association*, 111, 132–144.

Kiviet, J. F. and S. Kripfganz (2021): "Instrument approval by the Sargan test and its consequences for coefficient estimation," *Economics Letters*, 205.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2007): "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

Kling, J. R. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876.

Maasoumi, E. and P. C. Phillips (1982): "On the behavior of inconsistent instrumental variable estimators," *Journal of Econometrics*, 19, 183–201.

Masten, M. A. and A. Poirier (2021): "Salvaging Falsified Instrumental Variable Models," *Econometrica*, 89, 1449–1469.

Mikusheva, A. (2010): "Robust confidence sets in the presence of weak instruments," *Journal of Econometrics*, 157, 236–247.

Mikusheva, A. and L. Sun (2021): "Inference with Many Weak Instruments," *The Review of Economic Studies*, 89, 2663–2686.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Olea, J. L. M. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Sargan, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393.

Singh, R. and L. Sun (2021): "Automatic Kappa Weighting for Instrumental Variable Models of Complier Treatment Effects," *arXiv:1909.05244v5*.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stevenson, M. T. (2018): "Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes," *The Journal of Law, Economics, and Organization*, 34, 511–542.

Stock, J. and M. Yogo (2005): *Testing for Weak Instruments in Linear IV Regression*, New York: Cambridge University Press, 80–108.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2018): "On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments," *Journal of the American Statistical Association*, 114, 1339–1350.

Windmeijer, F., X. Liang, F. P. Hartwig, and J. Bowden (2021): "The confidence interval method for selecting valid instrumental variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 752–776.

A Figures and tables

A.1 Directed acyclic graphs

Figure 1: Linear IV model with endogenous IVs DAG, direct effect (top); Linear IV model with endogenous IVs DAG, indirect effect (bottom)

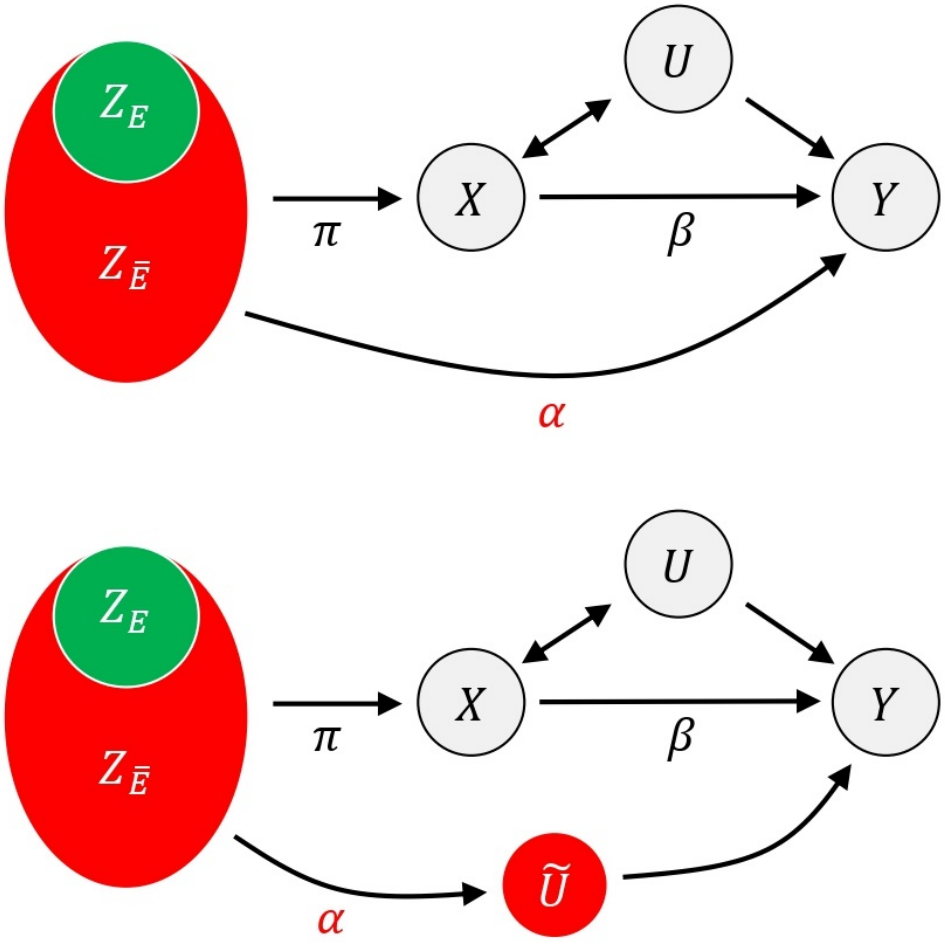
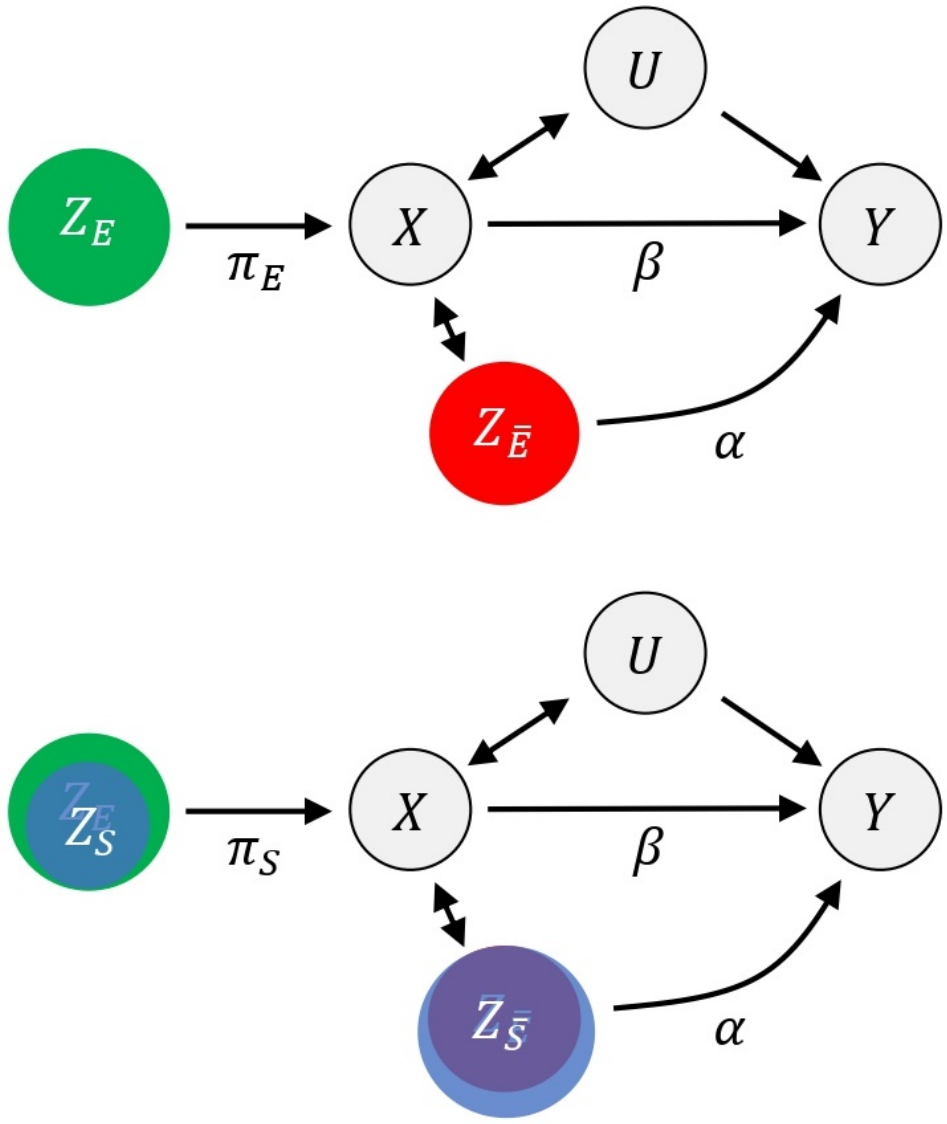


Figure 2: Linear IV model with endogenous IVs DAG, oracle model (top); Linear IV model with endogenous IVs DAG, valid model (bottom)



A.2 Simulation results

Table 1: 2SLS diagnostics by selection method, general setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.299	1.087	1.181	0.000	0.392	3.421
	AN	4.434	2.143	4.592	0.337	0.889	3.268
	Post-lasso	1.687	0.730	0.532	0.657	0.996	1.610
	Post-adalasso	1.819	0.726	0.527	0.562	0.705	2.576
	R_{EXO}	0.155	0.079	0.006	0.935	0.400	2.064
	R_{PMSE}	0.144	0.072	0.005	0.928	0.389	2.995
	R_{MSE}	0.172	0.087	0.008	0.816	0.373	2.566
	Oracle	0.144	0.072	0.005	0.954	0.399	1.000
n=4000	DN	0.201	1.112	1.237	0.000	0.133	3.422
	AN	0.130	2.013	4.053	0.278	0.227	2.720
	Post-lasso	1.925	0.929	0.863	0.602	0.241	2.284
	Post-adalasso	1.752	1.203	1.448	0.562	0.246	2.440
	R_{EXO}	0.046	0.023	0.001	0.948	0.126	2.058
	R_{PMSE}	0.043	0.022	0.000	0.951	0.125	2.998
	R_{MSE}	0.044	0.022	0.000	0.947	0.125	2.407
	Oracle	0.044	0.022	0.000	0.948	0.126	1.000

Table 2: 2SLS diagnostics by selection method, strong IVs setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.076	0.995	0.991	0.000	0.199	6.000
	AN	1.992	1.801	3.243	0.489	0.250	2.997
	Post-lasso	1.654	0.986	0.972	0.348	0.369	2.096
	Post-adalasso	0.133	0.992	0.984	0.172	0.209	4.303
	R_{EXO}	0.082	0.042	0.002	0.908	0.217	2.766
	R_{PMSE}	0.078	0.040	0.002	0.892	0.212	2.990
	R_{MSE}	0.083	0.042	0.002	0.843	0.208	3.127
	Oracle	0.074	0.037	0.001	0.940	0.211	3.000
n=4000	DN	0.025	1.001	1.002	0.000	0.063	6.000
	AN	1.997	1.944	3.778	0.458	0.109	2.999
	Post-lasso	0.041	1.001	1.003	0.151	0.064	4.163
	Post-adalasso	0.025	1.001	1.002	0.000	0.063	6.000
	R_{EXO}	0.026	0.013	0.000	0.919	0.068	2.868
	R_{PMSE}	0.023	0.012	0.000	0.934	0.068	2.998
	R_{MSE}	0.023	0.011	0.000	0.945	0.068	3.000
	Oracle	0.023	0.011	0.000	0.945	0.068	3.000

Table 3: 2SLS diagnostics by selection method, strong IVs favorable setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.146	1.328	1.764	0.000	0.368	6.000
	AN	0.064	0.032	0.001	0.939	0.176	3.946
	Post-lasso	0.079	0.040	0.002	0.865	0.185	3.539
	Post-adalasso	0.066	0.033	0.001	0.920	0.177	3.898
	R_{EXO}	0.069	0.035	0.001	0.911	0.178	3.771
	R_{PMSE}	0.064	0.032	0.001	0.935	0.176	3.986
	R_{MSE}	0.064	0.032	0.001	0.927	0.175	4.014
	Oracle	0.063	0.032	0.001	0.939	0.176	4.000
n=4000	DN	0.046	1.335	1.783	0.000	0.117	6.000
	AN	0.020	0.010	0.000	0.948	0.056	3.986
	Post-lasso	0.024	0.013	0.000	0.880	0.058	3.463
	Post-adalasso	0.026	0.015	0.000	0.862	0.066	3.078
	R_{EXO}	0.021	0.011	0.000	0.932	0.056	3.909
	R_{PMSE}	0.020	0.010	0.000	0.947	0.056	3.998
	R_{MSE}	0.020	0.010	0.000	0.949	0.056	4.000
	Oracle	0.020	0.010	0.000	0.949	0.056	4.000

Table 4: 2SLS diagnostics by selection method, exogenous IVs setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.091	0.046	0.002	0.944	0.263	3.417
	AN	0.091	0.046	0.002	0.933	0.258	5.923
	Post-lasso	0.116	0.063	0.004	0.947	0.384	1.230
	Post-adalasso	0.097	0.048	0.002	0.930	0.259	5.760
	R_{EXO}	0.118	0.058	0.003	0.927	0.275	3.099
	R_{PMSE}	0.091	0.045	0.002	0.940	0.257	5.982
	R_{MSE}	0.091	0.045	0.002	0.935	0.257	5.104
	Oracle	0.091	0.046	0.002	0.953	0.271	2.000
n=4000	DN	0.029	0.015	0.000	0.942	0.085	3.421
	AN	0.029	0.015	0.000	0.941	0.083	5.971
	Post-lasso	0.038	0.020	0.000	0.945	0.124	1.577
	Post-adalasso	0.032	0.016	0.000	0.929	0.084	5.460
	R_{EXO}	0.033	0.016	0.000	0.925	0.086	3.269
	R_{PMSE}	0.029	0.015	0.000	0.946	0.083	5.993
	R_{MSE}	0.029	0.015	0.000	0.946	0.083	5.806
	Oracle	0.030	0.015	0.000	0.942	0.086	2.000

Table 5: CLR diagnostics by selection method

		Diagnostics											
		General setting			Strong IV setting			Strong IV favorable setting			Exogenous IV setting		
		Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI
n=400	DN	0.00	1.51	1.00	0.00	0.24	1.00	0.00	1.04	1.00	0.95	0.27	1.00
	AN	0.46	0.71	0.56	0.47	0.29	1.00	0.95	0.18	1.00	0.94	0.27	1.00
	Post-lasso	0.65	0.65	0.62	0.34	0.38	1.00	0.90	0.19	1.00	0.96	0.39	0.96
	Post-adalasso	0.58	0.62	0.71	0.16	0.26	1.00	0.94	0.18	1.00	0.94	0.27	1.00
	R_{EXO}	0.94	0.41	0.94	0.93	0.22	1.00	0.93	0.18	1.00	0.93	0.27	0.95
	R_{PMSE}	0.94	0.41	1.00	0.93	0.22	1.00	0.95	0.18	1.00	0.95	0.27	1.00
	R_{MSE}	0.84	0.41	0.98	0.86	0.21	1.00	0.94	0.18	1.00	0.95	0.27	1.00
	Oracle	0.95	0.41	1.00	0.95	0.22	1.00	0.95	0.18	1.00	0.95	0.28	1.00
n=4000	DN	0.00	0.64	1.00	0.00	0.08	1.00	0.00	0.33	1.00	0.95	0.09	1.00
	AN	0.33	0.22	0.67	0.48	0.07	1.00	0.95	0.06	1.00	0.94	0.08	1.00
	Post-lasso	0.65	0.21	0.57	0.15	0.08	1.00	0.90	0.06	1.00	0.95	0.12	0.98
	Post-adalasso	0.61	0.22	0.53	0.00	0.08	1.00	0.88	0.07	1.00	0.93	0.08	1.00
	R_{EXO}	0.95	0.13	0.95	0.93	0.07	1.00	0.93	0.06	1.00	0.94	0.09	0.95
	R_{PMSE}	0.95	0.13	1.00	0.94	0.07	1.00	0.95	0.06	1.00	0.95	0.08	1.00
	R_{MSE}	0.95	0.13	0.99	0.95	0.07	1.00	0.95	0.06	1.00	0.95	0.08	1.00
	Oracle	0.95	0.13	1.00	0.95	0.07	1.00	0.95	0.06	1.00	0.95	0.09	1.00

A.3 Application figures and tables

Table 6: Judge descriptive statistics

Descriptive Statistic	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8
% cases allocated	0.065	0.039	0.163	0.170	0.101	0.166	0.125	0.170
% pre-trial detention	0.402	0.432	0.418	0.395	0.413	0.432	0.413	0.398
% pre-trial detention, cond. male	0.423	0.450	0.448	0.422	0.440	0.462	0.437	0.423
% pre-trial detention, cond. black	0.452	0.487	0.467	0.451	0.473	0.487	0.465	0.452
% pre-trial detention, cond. one prior	0.438	0.479	0.466	0.433	0.462	0.475	0.460	0.445
% pre-trial detention, cond. three priors	0.474	0.529	0.511	0.468	0.514	0.523	0.508	0.490
% pre-trial detention, cond. possess	0.213	0.255	0.182	0.160	0.167	0.186	0.197	0.219
% pre-trial detention, cond. agg assault	0.515	0.475	0.524	0.530	0.541	0.577	0.555	0.486
% pre-trial detention, cond. felony	0.568	0.566	0.589	0.583	0.604	0.609	0.588	0.545
% pre-trial detention, cond. misdemeanor	0.389	0.414	0.403	0.385	0.400	0.419	0.402	0.389

Table 7: Balance check: mean variable by judge and by pretrial detention status

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Pretrial detention	No pretrial detention
Prior # cases	4.820	4.660	4.890	4.880	4.990	4.830	4.920	4.910	3.900	6.290
Prior # felony charges	0.999	0.808	1.320	1.320	1.580	1.320	1.480	1.300	0.984	1.810
Prior # guilt	1.440	1.360	1.630	1.640	1.790	1.640	1.740	1.640	1.320	2.110
Felony	0.518	0.522	0.503	0.509	0.512	0.512	0.503	0.503	0.360	0.720
Misdemeanor	0.932	0.932	0.932	0.935	0.933	0.933	0.935	0.935	0.953	0.906
Summary offense	0.049	0.049	0.060	0.059	0.063	0.057	0.065	0.059	0.056	0.064
Felony type 1	0.142	0.149	0.141	0.144	0.137	0.142	0.137	0.138	0.056	0.262
Felony type 2	0.110	0.111	0.116	0.117	0.116	0.114	0.115	0.112	0.052	0.205
Felony type 3	0.223	0.218	0.202	0.204	0.194	0.203	0.195	0.201	0.126	0.312
Other felony	0.133	0.153	0.129	0.135	0.142	0.138	0.131	0.135	0.127	0.147
Misdemeanor type 1	0.365	0.357	0.364	0.364	0.355	0.360	0.360	0.356	0.260	0.504
Misdemeanor type 2	0.374	0.363	0.373	0.368	0.363	0.366	0.363	0.366	0.285	0.484
Misdemeanor type 3	0.078	0.079	0.081	0.078	0.076	0.079	0.085	0.078	0.075	0.085
Other misdemeanors	0.425	0.444	0.409	0.419	0.415	0.419	0.411	0.422	0.512	0.283
Robbery	0.078	0.076	0.070	0.077	0.068	0.076	0.072	0.070	0.021	0.148
Aggravated assault	0.085	0.079	0.092	0.090	0.096	0.091	0.091	0.090	0.072	0.117
Drug possession	0.156	0.133	0.133	0.137	0.118	0.137	0.134	0.136	0.186	0.062
Selling drugs	0.125	0.144	0.121	0.127	0.132	0.128	0.123	0.127	0.124	0.131
1st offense DUI	0.063	0.059	0.067	0.066	0.068	0.063	0.067	0.065	0.099	0.016
Guilt	0.450	0.452	0.491	0.489	0.523	0.495	0.505	0.492	0.492	0.493
Bail date	13,858	13,667	14,542	14,522	15,018	14,497	14,804	14,512	14,514	14,548
White	0.283	0.278	0.289	0.282	0.273	0.285	0.287	0.284	0.300	0.260
Black	0.604	0.586	0.569	0.576	0.568	0.576	0.575	0.577	0.524	0.651
Age	32.1	32.1	32.6	32.4	32.7	32.4	32.6	32.6	32.8	32.0
Male	0.830	0.831	0.829	0.836	0.828	0.837	0.829	0.831	0.795	0.885
One prior	0.759	0.753	0.761	0.761	0.768	0.759	0.764	0.762	0.704	0.844
Three priors	0.513	0.518	0.523	0.519	0.532	0.522	0.524	0.522	0.444	0.635
Waiting for another trial	0.638	0.638	0.636	0.643	0.646	0.638	0.640	0.638	0.567	0.744
Morning pretrial	0.326	0.309	0.355	0.343	0.363	0.350	0.364	0.351	0.359	0.337
Evening pretrial	0.344	0.352	0.321	0.331	0.336	0.328	0.332	0.333	0.335	0.326
Early morning pretrial	0.329	0.339	0.324	0.326	0.301	0.322	0.304	0.316	0.306	0.337
Weekend pretrial	0.029	0.039	0.038	0.034	0.029	0.031	0.033	0.035	0.034	0.033
Day of the year	168	142	199	200	221	197	217	199	198	199
Monday	0.577	1	0.234	0.238	0	0.258	0	0.243	0.242	0.237
Tuesday	0.423	0	0.169	0.174	0	0.177	0.219	0.176	0.173	0.170
Wednesday	0	0	0.161	0.168	0.252	0.151	0.213	0.160	0.165	0.152
Thursday	0	0	0.144	0.146	0.254	0.148	0.193	0.150	0.150	0.145
Friday	0	0	0.150	0.137	0.252	0.135	0.196	0.143	0.145	0.144
Saturday	0	0	0.142	0.137	0.242	0.131	0.178	0.127	0.126	0.152

Table 8: Estimates of judge fixed effects on pre-trial detention, robust standard errors

	(1)	(2)	(3)
Judge 1	-0.0322^{****}	-0.0306^{****}	-0.0305^{****}
	(0.0056)	(0.0052)	(0.0050)
Judge 3	-0.0232^{****}	-0.0211^{****}	-0.0189^{****}
	(0.0051)	(0.0047)	(0.0046)
Judge 4	-0.0455^{****}	-0.0452^{****}	-0.0440^{****}
	(0.0051)	(0.0047)	(0.0045)
Judge 5	-0.0326^{****}	-0.0311^{****}	-0.0296^{****}
	(0.0056)	(0.0052)	(0.0050)
Judge 6	-0.0086^{*}	-0.0080^{*}	-0.0071
	(0.0051)	(0.0047)	(0.0046)
Judge 7	-0.0298^{****}	-0.0272^{****}	-0.0250^{****}
	(0.0055)	(0.0050)	(0.0049)
Judge 8	-0.0419^{****}	-0.0373^{****}	-0.0358^{****}
	(0.0051)	(0.0047)	(0.0046)
Time effects	Yes	Yes	Yes
Case characteristics	-	Yes	Yes
Offender characteristics	-	-	Yes
F statistic	35.22^{****}	40.39^{****}	43.06^{****}
J statistic	8.43	10.88^{*}	12.16^{*}

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 9: Estimates of effect of pre-trial detention on guilt, robust standard errors

		(1)	(2)	(3)
OLS	Est	0.0002	0.0563****	0.0289****
	Sd	(0.0018)	(0.0019)	(0.0019)
2SLS	Est	0.1509	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS AN	Est	0.1509**	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS DN	Est	0.1509**	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS Post-lasso	Est	0.2549**	0.2552**	0.2594**
	Sd	0.1166	0.1108	0.1127
	Judge IV	{4}	{4}	{4}
2SLS Post-adalasso	Est	0.2117***	0.2018***	0.5206
	Sd	0.07552	0.06505	0.7508
	Judge IV	{4;6}	{3;4;5;6;7;8}	{6}
2SLS R_{EXO}	Est	0.3832**	0.5025***	0.2595**
	Sd	0.1504	0.1606	0.1126
	Judge IV	{6;7}	{6;7}	{4;8}
2SLS R_{PMSE}	Est	0.2549**	0.2556**	0.2594**
	Sd	0.1166	0.1108	0.1127
	Judge IV	{4}	{4;8}	{4}
2SLS R_{MSE}	Est	0.06212	0.1190	-0.01643
	Sd	0.1086	0.1171	0.1477
	Judge IV	{1;3;7;8}	{1;3;7;8}	{1;6}
Time effects		Yes	Yes	Yes
Case characteristics		-	Yes	Yes
Demographics and other		-	-	Yes

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 10: Diagnostics post selection when estimating the effect of pre-trial detention on guilt, cluster robust standard errors

Selection method		(1)	(2)	(3)
No selection	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
AN	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
DN	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
Lasso	F statistic	80.2850****	97.8957****	99.6765****
	J statistic	0.0000	0.0000	0.0000
Adalasso	F statistic	93.9719****	46.5946****	2.5798
	J statistic	0.2422	6.9497	0.0000
R_{EXO}	F statistic	25.9478****	26.3822****	49.9060****
	J statistic	0.0416	0.0050	0.0019
R_{PMSE}	F statistic	80.2850****	48.9627****	99.6765****
	J statistic	0.0000	0.0439	0.0000
R_{MSE}	F statistic	21.8693****	21.2626****	27.8508****
	J statistic	7.1293*	9.0086**	0.6231
Time effects		Yes	Yes	Yes
Case characteristics		-	Yes	Yes
Demographics and other		-	-	Yes

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 11: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (1)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.1205	-0.8127	0.5811	19.3040	-0.02174	-3.3776	0.4673
	2	0.1205	-	0.4841*	0.2549**	0.3633**	0.5105	0.4096**	0.2008
	3	-0.8127	0.4841*	-	0.01663	0.0668	0.4684**	0.1470	-0.1495
	4	0.5811	0.2549**	0.01663	-	-0.02023	0.1952**	-0.03786	0.8918
	5	19.3040	0.3633**	0.0668	-0.02023	-	0.3105**	-0.1168	-0.3692
	6	-0.02174	0.5105	0.4684**	0.1952**	0.3105**	-	0.3685**	0.1208
	7	-3.3776	0.4096**	0.1470	-0.03786	-0.1168	0.3685**	-	-0.3096
	8	0.4673	0.2008	-0.1495	0.8918	-0.3692	0.1208	-0.3096	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 12: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (2)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.05449	-0.7387	0.6742**	21.3701	-0.0888	-3.4442	1.0862
	2	0.05449	-	0.4113*	0.2552**	0.4229**	0.4563	0.4928**	0.2396*
	3	-0.7387	0.4113*	-	0.1188	0.4473	0.3835*	0.7754	0.01574
	4	0.6742**	0.2552**	0.1188	-	-0.1147	0.2117***	-0.1025	0.3281
	5	21.3701	0.4229**	0.4473	-0.1147	-	0.4112***	-0.05897	-0.6879
	6	-0.0888	0.4563	0.3835*	0.2117***	0.4112***	-	0.5081***	0.1800*
	7	-3.4442	0.4928**	0.7754	-0.1025	-0.05897	0.5081***	-	-0.4422
	8	1.0862	0.2396*	0.01574	0.3281	-0.6879	0.1800*	-0.4422	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 13: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (3)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.05902	-0.5675	0.7104**	-12.5002	-0.08053	-2.2185	1.3849
	2	0.05902	-	0.4449	0.2594**	0.4376**	0.5206	0.5577**	0.2560*
	3	-0.5675	0.4449	-	0.1202	0.4248	0.3995	0.9041	0.04559
	4	0.7104**	0.2594**	0.1202	-	-0.1062	0.2093***	-0.1330	0.2740
	5	-12.5002	0.4376**	0.4248	-0.1062	-	0.4115***	-0.2176	-0.6094
	6	-0.08053	0.5206	0.3995	0.2093***	0.4115***	-	0.5724***	0.1908*
	7	-2.2185	0.5577**	0.9041	-0.1330	-0.2176	0.5724***	-	-0.4430
	8	1.3849	0.2560*	0.04559	0.2740	-0.6094	0.1908*	-0.4430	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

B Application appendix

In this section, I justify the choice of "excluding" judge 2 from the set of eight potential judge dummy IVs and guess which judge dummy is possibly endogenous. First, notice that by projecting out the exogenous regressors except the intercept the model can be written as

$$guilt_i = \gamma_0 + predet_i\beta + \sum_{j\in\overline{E}} \alpha_j 1\{judge_i = j\} + u_i$$

$$predet_i = \delta_0 + \sum_{j=1}^{8} \pi_j 1\{judge_i = j\} + v_i$$

Then let for $j = 1,\dots,8$ $p_j = \mathbb{P}(judge_i = j)$, $\overline{predet} = \mathbb{E}(predet_i)$ and $\overline{guilt} = \mathbb{E}(guilt_i)$ so that the demeaned variables can be written for $j = 1,\dots,8$ as $D_{ij} = 1\{judge_i = j\} - p_j$, $\tilde{predet}_i = predet_i - \overline{predet}$ and $\tilde{guilt}_i = guilt_i - \overline{guilt}$. The intercept can then be projected out so that the model can be rewritten as

$$\tilde{guilt}_i = \tilde{predet}_i\beta + \sum_{j\in\overline{E}} \alpha_j D_{ij} + \tilde{u}_i$$

$$\tilde{predet}_i = \sum_{j=1}^{8} \pi_j D_{ij} + \tilde{v}_i$$

Thus, because $\sum_{j=1}^{8} D_{ij} = 0$ there is a multicollinearity problem in the first stage and some judge variable must be removed.

Then what if, for instance, judge 8 is excluded and not used either as an IV or as a control variable? Can it still enter the structural equation? If judge 8 is excluded and enters the structural equation, then it is actually $D_{i8} = 1\{judge_i = 8\} - p_8 = \sum_{j=1}^{7}(p_j - 1\{judge_i = j\}) = \sum_{j=1}^{7} D_{ij}$ which enters the structural equation. Hence, affirming that excluded judge 8 enters the structural equation implies that there is at least one other judge dummy variable which is endogenous which is the usual case considered in the paper. On the other hand, affirming that judge 8 doesn't enter the structural equation, and thus doesn't imply that all other judge dummy variables are exogenous hence it is also the usual case. Consequently, endogeneity or exogeneity of the excluded variable is entirely tied to endogeneity or exogeneity of the rest of the judge dummy variables which can be dealt with IV selection methods.

Still, in order to limit as much as possible the overall amount of endogeneity I exclude the judge which has the lowest amount of cases, ie judge 2 with $\hat{p}_2 = 0.039$, see 6 in appendix A.3. Indeed, if judge 2 is excluded but still enters the structural equation through the sum of the other dummies, endogeneity is limited compared to when excluding other judges. When the model is instrumented by subset S with $\overline{S}$ projected out

$$\tilde{guilt}_i = \tilde{predet}_i\beta + \tilde{u}_{iS}, \qquad \tilde{predet}_i = \sum_{j\in S} \pi_j D_{ij} + v_i$$

where the error is $\tilde{u}_{iS} = \alpha_2 D_{i2} + \sum_{j \in \overline{E}} \alpha_j D_{ij} \tilde{u}_i$. Thus, for any S part of its level of endogeneity is determined by $\mathbb{E}\left(\sum_{j \in S} D_{ij} D_{i2}\right) = -\mathbb{P}(judge_i = 2) \sum_{j \in S} \mathbb{P}(judge_i = j)$ which is small because $\mathbb{P}(judge_i = 2)$ is the smallest out of all judge propensity to take a case.

In addition, as mentioned before, excluding judge j implies that all other judge effects are relative to judge j in the first stage. This means that the IVs which are individually insignificant in the first stage depend on the judge which is excluded. Thus, the excluded judge determines which IV subsets are weak, this is quite problematic because IV selection methods are sensitive to weak IVs, including the ones developed in this paper to some extent. For this reason excluding judge 2 is a great choice because judge 2 has the largest propensity to send someone to pre-trial detention, thus when excluding judge 2 other judge dummies are more likely to be individually significant.

In tables 11, 12, and 13 are 2SLS estimators computed for specification (1), (2) and (3) respectively, where each column correspond to which judge was excluded, and where each row corresponds to which judge was used as the sole IV to construct 2SLS (other judge dummies act as controls). First, notice that the tables are symmetric, this is due to the fact that if in the model six judge dummies and an intercept are controlled for then the two last judge dummy variables are equal to each other hence excluding one or the other does not change the estimator. Second, note that across specifications when judge 2 and to a lesser extent judge 8 are excluded (or used as the IV) the 2SLS estimators are much more likely to be significant, this is because they are the harshest judges and relative to them other judge dummies have a significant effect on pre-trial detention. Third, notice that across specifications some entries are extremes, sometimes negative, especially when judge 5 or judge 7 are excluded (or used as the IV). This is either due to the fact that being imprisoned before a trial is taken as a signal that the offender is not guilty, which is implausible, or this is due to these judge dummy variables being endogenous and producing a negative bias. Indeed, if the true model is D_{ij} is used as the sole IV but enters the structural equation as in

$$\tilde{guilt}_i = \tilde{predet}_i \beta + \alpha_j D_{ij} + \tilde{u}_i, \qquad \tilde{predet}_i = \pi_j D_{ij} + \tilde{v}_i$$

This can generate a negative bias when $sign(\alpha_j) \neq sign(\pi_j)$. If a pre-trial judge is too lenient with a negative π_j then the judge which supervises the trial may compensate and be harsher with a positive α_j. Fourth the differences between the individual judge IV estimators in tables 11, 12 and 13 could also be due (in part) to heterogeneity in β or to endogeneity of all judge IVs, although both are unlikely.

To conclude, there are very good reasons to exclude judge 2.

C Main theorems proofs

The proofs of the main asymptotic results are divided in three parts. First, the limit in distribution of all the risks, risk estimators and the limit of their expectation are found, then I

rank IV sets in accordance to the limits of each risk. Second, I prove that the feasible risk estimators and the risks have the same limit in probability. Third, I combine these results to prove the efficiency and the consistency of selection via risk estimator minimization. Note that, in order to simplify calculations throughout this proof, instead of considering $\hat{R}_{EXO}$ and R_{EXO} I consider $n_b\hat{R}_{EXO}$ and $\tilde{R}_{EXO}$ (which is defined below).

Before writing the proofs, some notations and conventions are introduced. Unless specified, all limits are taken with respect to n. Additionally, we denote by the expression $X = o_P(n^a)$ a random variable or statistic X which is asymptotically degenerate of order n^a, ie $X = o_P(n^a) \Leftrightarrow \forall e > 0 \ \mathbb{P}(|X|n^{-a} > e) \underset{n\to\infty}{\to} 0$, and denote by $X = O_p(n^a)$ a random variable which is (bounded in probability) of order n^a, ie $\forall e > 0 \ \exists M > 0, \exists N : \forall n \geqslant N \ \mathbb{P}(|X|n^{-a} > M) < e$. The usual properties of o_P and O_P random variables are used throughout these proofs. In adition, *plim* X denotes the limit in probability of a random variable or statistic X whereas *dlim* X denotes its limit in distribution.

In all the proofs, denote the difference between the subset IV estimator and the true parameter as $\hat{C}_S \equiv \hat{\beta}_S - \beta = \frac{x'P_{z_S}u_S}{x'P_{z_S}x}$ for any $S \in \mathcal{S}$. Then, I define the following risks for IV subsets based on exogeneity as

$$\tilde{R}_{EXO}(S) = \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)'z_S^*\Sigma_S z_S^{*'}(y^* - x^*\hat{\beta}_S)\right)\right)$$

$$\overline{R}_{EXO}(S) = \begin{cases} R_{EXO}(S) & \text{if } b_S \leqslant 1/2 \\ \tilde{R}_{EXO}(S) & \text{if } b_S \geqslant 1/2 \end{cases}$$

In the later parts of the proof, R_{EXO} and $\tilde{R}_{EXO}$ are used altenatively to rank IV subsets. Finally, I define the out-of-sample risk estimators

$$\hat{R}_{EXO,o}(S) = \frac{1}{n^2}(y^* - x^*\hat{\beta}_S)'z_S^*\Sigma_S z_S^{*'}(y^* - x^*\hat{\beta}_S)$$

$$\hat{R}_{PMSE,o}(S) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - z_{iS}^{*'}\pi_S\hat{\beta}_S)^2$$

$$\hat{R}_{MSE,o}(S) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - x_i^*\hat{\beta}_S)^2$$

where $\hat{\beta}_S$ was computed from the original sample $(w_i)_{i=1}^n$ and the sample $(w_i^*)_{i=1}^n$ has the same DGP but is independent of $(w_i)_{i=1}^n$.

C.1 Technical lemmas: expected risk estimators asymptotics and ranking

Lemma 3.1

Under Assumption A, for any $S \in \mathcal{S}$

- *If* $a_S \in [0; 1/2)$ *then*
  - *If* $b_S > a_S$ *then plim* $\hat{C}_S =$ *plim* $\hat{C}_S^2 = 0$, *in addition if* $b_S \geqslant 1/2$ *then* $n^{1/2-a_S}\hat{C}_S = O_P(1)$
  - *If* $b_S = a_S$ *then plim* $\hat{C}_S = \frac{\kappa_S'\delta_S}{\kappa_S'\Sigma_S\kappa_S}$
  - *If* $b_S < a_S$ *then* $\hat{C}_S$ *has no limit but* $n^{-a_S}\hat{C}_S = O_P(1)$ *and if* $b_S = 0$ *then plim* $n^{-a_S}\hat{C}_S = \frac{\kappa_S'\delta_S}{\kappa_S\Sigma_S\kappa_S}$
- *If* $a_S = 1/2$ *then*
  - *If* $b_S > 1/2$ *then dlim* $\hat{C}_S = \frac{\lambda_{vS}'\lambda_u}{||\lambda_{vS}||^2}$
  - *If* $b_S = 1/2$ *then dlim* $\hat{C}_S = \frac{\lambda_{vS}'(\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_{vS}||^2}$
  - *If* $b_S < 1/2$ *then* $\hat{C}_S$ *has no limit but* $n^{-1/2}\hat{C}_S = O_P(1)$ *and if* $b_S = 0$ *then dlim* $n^{-1/2}\hat{C}_S = \frac{\lambda_{vS}'\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}$
- *If* $a_S > 1/2$ *then*
  - *If* $b_S > 1/2$ *then dlim* $\hat{C}_S = \frac{\lambda_v'\lambda_u}{||\lambda_v||^2}$
  - *If* $b_S = 1/2$ *then dlim* $\hat{C}_S = \frac{\lambda_v'(\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_v||^2}$
  - *If* $b_S < 1/2$ *then* $\hat{C}_S$ *has no limit but* $n^{-1/2}\hat{C}_S = O_P(1)$ *and if* $b_S = 0$ *then dlim* $n^{-1/2}\hat{C}_S = \frac{\lambda_v'\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}$

where (λ_u, λ_v) *is Gaussian and* $\lambda_{vS} = \lambda_v + \Sigma_S^{1/2}\kappa_S$. *Furthermore, if* $b_S \geqslant 1/2$ *then* $\hat{C}_S = O_P(1)$.

Proof. Take note of the following decomposition of $\hat{C}_S$

$$\begin{aligned}
\hat{C}_S = \frac{x'P_{z_S}u}{x'P_{z_S}x} &= \frac{\pi_S'z_S'z_{\bar{E}}'\alpha + v'P_{z_S}z_{\bar{E}}\alpha + v'P_{z_S}u + \pi_S'z_S'u}{\pi_S'z_S'z_S\pi_S + 2\pi_S'z_Sv + v'P_{z_S}v} \\
&= \frac{n^{1-a_S-b_S}\kappa_S'\delta_S + n^{1-b_S}v'z_S(z_S'z_S)^{-1}\delta_S + v'P_{z_S}u + n^{-a_S}\kappa_S'z_S'u}{n^{-2a_S}\kappa_S'z_S'z_S\kappa_S + 2n^{-a_S}\kappa_S'z_Sv + v'P_{z_S}v} \\
&\quad + \frac{n^{-a_S}\kappa_S'z_S'z_{\bar{E}}\alpha - n^{1-a_S-b_S}\kappa_S'\delta_S + v'P_{z_S}z_{\bar{E}}\alpha - n^{1-b_S}v'z_S(z_S'z_S)^{-1}\delta_S}{n^{-2a_S}\kappa_S'z_S'z_S\kappa_S + 2n^{-a_S}\kappa_S'z_Sv + v'P_{z_S}v} \\
&= \frac{n^{1-a_S-b_S}\kappa_S'\delta_S + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}\delta_S + v'P_{z_S}u + n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_S'u}{n^{1-2a_S}\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v} \\
&\quad + \frac{n^{1-a_S-b_S}\kappa_S'(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S) + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S)}{n^{1-2a_S}\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v}
\end{aligned}$$

Then by assumption A the (weak) law of large numbers (LLN), the central limit theorem (CLT) and the continuous mapping theorem (CMT) can be applied to obtain the limits of the components of $\hat{C}_S$

$$\begin{aligned}
&\frac{1}{\sqrt{n}} v' z_S \xrightarrow{d} \mathcal{N}(0, \sigma_v^2 \Sigma_S) \\
&\frac{1}{\sqrt{n}} u' z_S \xrightarrow{d} \mathcal{N}(0, \sigma_u^2 \Sigma_S) \\
&\frac{1}{n} z_S' z_S \xrightarrow{\mathbb{P}} \Sigma_S \\
&n^{b_S - 1} z_S' z_{\bar{E}} \alpha \xrightarrow{\mathbb{P}} \delta_S
\end{aligned}$$

Then denote

$$\begin{aligned}
\lambda_u &\equiv dlim(n^{-1/2} z_S' u \Sigma_S^{-1/2}) \sim \mathcal{N}(0, \sigma_u^2 I_s) \\
\lambda_v &\equiv dlim(n^{-1/2} z_S' v \Sigma_S^{-1/2}) \sim \mathcal{N}(0, \sigma_v^2 I_s) \\
\forall S,\ \lambda_{vS} &\equiv dlim(n^{-1/2} z_S' v \Sigma_S^{-1/2} + \Sigma_S^{1/2} \kappa_S) \equiv \lambda_v + \Sigma_S^{1/2} \kappa_S \sim \mathcal{N}(\Sigma_S^{1/2} \kappa_S, \sigma_v^2 I_s)
\end{aligned}$$

and note that for any $j = 1, \ldots, s$ $\mathbb{E}(\lambda_{u,j} \lambda_{v,j}) = \mathbb{E}(\lambda_{u,j} \lambda_{vS,j}) = \rho$ but $\mathbb{E}(\lambda_{u,j} \lambda_{v,j'}) = \mathbb{E}(\lambda_{u,j} \lambda_{vS,j'}) = 0$ if $j' \neq j$. Thus, by properties of Gaussian vectors for any S

$$\lambda_u = \frac{\rho}{\sigma_v^2} \lambda_v + \varepsilon = \frac{\rho}{\sigma_v^2} \lambda_{vS} - \frac{\rho}{\sigma_v^2} \Sigma_S^{1/2} \kappa_S + \varepsilon$$

where ε_j is independent of $\lambda_{v,j}$ and $\varepsilon \sim \mathcal{N}\left(0, (\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}) I_s\right)$.

Consequently, with a slight abuse of notations $\hat{C}_S$ can be rewritten in terms of O_P orders

$$\begin{aligned}
\hat{C}_S = &\frac{O_P(n^{1-a_S-b_S}) + O_P(n^{1/2-b_S}) + O_P(1) + O_P(n^{1/2-a_S})}{O_P(n^{1-2a_S}) + O_P(n^{1/2-a_S}) + O_P(1)} \\
&+ \frac{O_P(n^{1-a_S-b_S}) o_P(1) + O_P(n^{1/2-b_S}) o_P(1)}{O_P(n^{1-2a_S}) + O_P(n^{1/2-a_S}) + O_P(1)}
\end{aligned}$$

Thus, (a_S, b_S) determines the asymptotic behavior of $\hat{C}_S$ and there are six cases

- When $a_S < 1/2$ (S is a semi-strong IV subset) $\hat{C}_S$ can be rewritten as

$$\hat{C}_S = \hat{C}_S \frac{n^{2a_S-1}}{n^{2a_S-1}} = \frac{n^{a_S-b_S}\kappa_S'\delta_S + n^{2a_S-1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}\delta_S + n^{2a_S-1}v'P_{z_S}u + n^{a_S-1/2}\kappa_S'\frac{1}{\sqrt{n}}z_S'u}{\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{a_S-1/2}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + n^{2a_S-1}v'P_{z_S}v}$$
$$+ \frac{n^{a_S-b_S}\kappa_S'(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S) + n^{2a_S-1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S)}{\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{a_S-1/2}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + n^{2a_S-1}v'P_{z_S}v}$$
$$= \frac{O_P(n^{a_S-b_S}) + O_P(n^{2a_S-1/2-b_S}) + O_P(n^{2a_S-1}) + O_P(n^{a_S-1/2})}{O_P(1) + O_P(n^{a_S-1/2}) + O_P(n^{2a_S-1})}$$
$$+ \frac{O_P(n^{a_S-b_S})o_P(1) + O_P(n^{2a_S-1/2-b_S})o_P(1)}{O_P(1) + O_P(n^{a_S-1/2}) + O_P(n^{2a_S-1})}$$

But $O_P(1) + O_P(n^{a_S-1/2}) + O_P(n^{2a_S-1}) = O_P(1)$, $O_P(n^{2a_S-1}) = o_P(1)$, and $O_P(n^{a_S-1/2}) = o_P(1)$ because $a_S < 1/2$ thus

$$\hat{C}_S = O_P(n^{a_S-b_S}) + O_P(n^{2a_S-1/2-b_S}) + o_P(1)$$

Then there are three cases

- When $b_S > a_S$ (S is an exogenous IV subset) then $O_P(n^{a_S-b_S}) = o_P(1)$ and $O_P(n^{a_S-1/2+a_S-b_S}) = o_P(1)$ because $a_S < 1/2$ so that $\hat{C}_S = o_P(1)$ and $\hat{\beta}_S$ is consistent. In addition, when $b_S \geqslant 1/2$ then $n^{1/2-a_S}\hat{C}_S = O_P(1)$ so that if $b_S > 1/2$ then $dlim\ n^{1/2-a_S}\hat{C}_S = \frac{\kappa_S'\Sigma_S^{1/2}\lambda_u}{\kappa_S'\Sigma_S\kappa_S}$ and

$$\mathbb{E}(\ dlim\ n^{1/2-a_S}\hat{C}_S) = 0, \qquad \mathbb{E}\left(\ dlim\left(n^{1/2-a_S}\hat{C}_S\right)^2\right) = \frac{\sigma_u^2}{\kappa_S'\Sigma_S\kappa_S}$$

  and so that if $b_S = 1/2$ then $dlim\ n^{1/2-a_S}\hat{C}_S = \frac{\kappa_S'\Sigma_S^{1/2}(\lambda_u + \Sigma_S^{-1/2}\delta_S)}{\kappa_S'\Sigma_S\kappa_S}$ and

$$\mathbb{E}(\ dlim\ n^{1/2-a_S}\hat{C}_S) = \frac{\kappa_S'\delta_S}{\kappa_S'\Sigma_S\kappa_S}, \qquad \mathbb{E}\left(\ dlim\left(n^{1/2-a_S}\hat{C}_S\right)^2\right) = \frac{\sigma_u^2 + (\kappa_S'\delta_S)^2}{\kappa_S'\Sigma_S\kappa_S}$$

- When $b_S = a_S$ (S is a locally endogenous IV subset) then $O_P(n^{a_S-b_S}) = O_P(1)$ and $O_P(n^{a_S-1/2+a_S-b_S}) = o_P(1)$ because $a_S = b_S < 1/2$ so that $\hat{C}_S = O_P(1)$ and $\hat{\beta}_S$ is inconsistent. More precisely, when $b_S = a_S < 1/2$ the limit in probability of $\hat{C}_S$ is $\frac{\kappa_S'\delta_S}{\kappa_S'\Sigma_S\kappa_S}$ thus $\hat{\beta}_S$ is inconsistent unless $\kappa_S'\delta_S = 0$.

- When $b_S < a_S$ (S is an endogenous IV subset) then $\hat{C}_S$ has no limit and $\hat{\beta}_S$ is inconsistent. But $n^{-a_S}\hat{C}_S = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{if } b_S = 0 \end{cases}$ and more precisely if $b_S = 0$ then $plim\ n^{-a_S}\hat{C}_S = \frac{\kappa_S'\delta_S}{\kappa_S'\Sigma_S\kappa_S}$.

- When $a_S = 1/2$ (S is a weak IV subset) $\hat{C}_S$ can be rewritten as

$$\begin{aligned}\hat{C}_S &= \frac{n^{1/2-b_S}\kappa_S'\delta_S + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}\delta_S + v'P_{z_S}u + \kappa_S'\frac{1}{\sqrt{n}}z_S'u}{\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v} \\ &\quad + \frac{n^{1/2-b_S}\kappa_S'(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S) + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S)}{\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v} \\ &= O_P(n^{1/2-b_S}) + O_P(1) + o_P(1)\end{aligned}$$

  Then there are three cases

  - When $b_S > 1/2$ (S is an exogenous IV subset) then $O_P(n^{1/2-b_S}) = o_P(1)$ so that $\hat{C}_S = O_P(1)$ and $\hat{\beta}_S$ is inconsistent. More precisely, when $b_S > a_S = 1/2$ the limit in distribution of $\hat{C}_S$ is

$$dlim\ \hat{C}_S = \frac{(\lambda_v + \Sigma_S^{1/2}\kappa_S)'\lambda_u}{(\lambda_v + \Sigma_S^{1/2}\kappa_S)'(\lambda_v + \Sigma_S^{1/2}\kappa_S)} = \frac{\lambda_{vS}'\lambda_u}{||\lambda_{vS}||^2}$$

    In addition, note that

$$\begin{aligned}dlim\ \hat{C}_S &= \frac{\rho}{\sigma_v^2} + \frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2}\frac{\lambda_{vS}}{||\lambda_{vS}||^2} \\ dlim\ \hat{C}_S^2 &= \frac{\rho^2}{\sigma_v^4} + \frac{\varepsilon'P_{\lambda_{vS}}\varepsilon}{||\lambda_{vS}||^2} + \frac{\rho^2}{\sigma_v^4}\kappa_S'\Sigma_S^{1/2}\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\Sigma_S^{1/2}\kappa_S + \\ &\quad + 2\frac{\rho}{\sigma_v^2}\frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} - 2\frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2}\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\varepsilon - 2\frac{\rho^2}{\sigma_v^4}\kappa_S'\Sigma_S^{1/2}\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\end{aligned}$$

$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S) = \frac{\rho}{\sigma_v^2}\left(1 - \kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\right)\right)$$

$$\begin{aligned}\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S^2) &= \frac{\rho^2}{\sigma_v^4}\left(1 - 2\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\right) + \kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\right)\Sigma_S^{1/2}\kappa_S\right) \\ &\quad + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)\mathbb{E}(||\lambda_{vS}||^{-2}) \\ &= \frac{\rho^2}{\sigma_v^4}\mathbb{E}\left(\left(1 - \frac{\lambda_{vS}'}{||\lambda_{vS}||^2}\Sigma_S^{1/2}\kappa_S\right)^2\right) + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)(\sigma_v^2(s-2))^{-1}\end{aligned}$$

– When $b_S = 1/2$ (S is a locally endogenous IV subset) then $\hat{C}_S = O_P(1)$ and $\hat{\beta}_S$ is inconsistent. More precisely, when $b_S = a_S = 1/2$ the limit in distribution of $\hat{C}_S$ is

$$dlim\ \hat{C}_S = \frac{\kappa_S'\delta_S + \lambda_v'\Sigma_S^{-1/2}\delta_S + (\lambda_v + \Sigma_S^{1/2}\kappa_S)'\lambda_u}{(\lambda_v + \Sigma_S^{1/2}\kappa_S)'(\lambda_v + \Sigma_S^{1/2}\kappa_S)} = \frac{\lambda_{vS}'(\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_{vS}||^2}$$

In addition, note that

$$\begin{aligned}dlim\ \hat{C}_S &= \frac{\rho}{\sigma_v^2} + \frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} + (\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\frac{\lambda_{vS}}{||\lambda_{vS}||^2} \\ dlim\ \hat{C}_S^2 &= \frac{\rho^2}{\sigma_v^4} + \frac{\varepsilon' P_{\lambda_{vS}}\varepsilon}{||\lambda_{vS}||^2} + (\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}(\Sigma_S^{-1/2}\delta_S - \frac{\rho}{\sigma_v^2}\Sigma_S^{1/2}\kappa_S) \\ &\quad + 2\frac{\rho}{\sigma_v^2}\frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} + 2(\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\varepsilon + 2\frac{\rho}{\sigma_v^2}(\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\end{aligned}$$

$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S) = \frac{\rho}{\sigma_v^2} + \left(\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2}\right)\mathbb{E}\left(\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\right)$$

$$\begin{aligned}\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S^2) &= \frac{\rho^2}{\sigma_v^4} + 2\frac{\rho}{\sigma_v^2}(\delta_S'\Sigma_S^{-1/2} - \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\mathbb{E}\left(\frac{\lambda_{vS}}{||\lambda_{vS}||^2}\right) \\ &\quad + (\delta_S'\Sigma_S^{-1/2} + \frac{\rho}{\sigma_v^2}\kappa_S'\Sigma_S^{1/2})\mathbb{E}\left(\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\right)(\Sigma_S^{-1/2}\delta_S + \frac{\rho}{\sigma_v^2}\Sigma_S^{1/2}\kappa_S) \\ &\quad + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)\mathbb{E}(||\lambda_{vS}||^{-2}) \\ &= \mathbb{E}\left(\left(\frac{\rho}{\sigma_v^2} + \frac{\lambda_{vS}'}{||\lambda_{vS}||^2}(\Sigma_S^{-1/2}\delta_S - \frac{\rho}{\sigma_v^2}\Sigma_S^{1/2}\kappa_S\right)^2\right) + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)(\sigma_v^2(s-2))^{-1}\end{aligned}$$

– When $b_S < 1/2$ (S is an endogenous IV subset) then $\hat{C}_S$ has no limit and $\hat{\beta}_S$ is

inconsistent. But $n^{-1/2}\hat{C}_S = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{if } b_S = 0 \end{cases}$ and more precisely if $b_S = 0$ then $dlim\ n^{-1/2}\hat{C}_S = \frac{\lambda_{vS}\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}$.

- When $a_S > 1/2$ (S is a very weak IV subset) recall that $\hat{C}_S$ can be rewritten as

$$\begin{aligned}
\hat{C}_S &= \frac{n^{1-a_S-b_S}\kappa_S'\delta_S + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}\delta_S + v'P_{z_S}u + n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_S'u}{n^{1-2a_S}\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v} \\
&\quad + \frac{n^{1-a_S-b_S}\kappa_S'(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S) + n^{1/2-b_S}\frac{1}{\sqrt{n}}v'z_S(\frac{1}{n}z_S'z_S)^{-1}(n^{b_S-1}z_S'z_{\bar{E}}\alpha - \delta_S)}{n^{1-2a_S}\kappa_S'\frac{1}{n}z_S'z_S\kappa_S + 2n^{1/2-a_S}\kappa_S'\frac{1}{\sqrt{n}}z_Sv + v'P_{z_S}v} \\
&= \frac{O_P(n^{1-a_S-b_S}) + O_P(n^{1/2-b_S}) + O_P(1) + O_P(n^{1/2-a_S})}{O_P(n^{1-2a_S}) + O_P(n^{1/2-a_S}) + O_P(1)} \\
&\quad + \frac{O_P(n^{1-a_S-b_S})o_P(1) + O_P(n^{1/2-b_S})o_P(1)}{O_P(n^{1-2a_S}) + O_P(n^{1/2-a_S}) + O_P(1)}
\end{aligned}$$

But $O_P(n^{1/2-a_S}) = o_P(1)$ and $O_P(n^{1-2a_S}) = o_P(1)$ because $a_S > 1/2$ thus

$$\hat{C}_S = O_P(n^{1-a_S-b_S}) + O_P(n^{1/2-b_S}) + O_P(1) + o_P(1)$$

Then there are three cases

  - When $b_S > 1/2$ (S is an exogenous IV subset) then $O_P(n^{1/2-b_S}) = o_P(1)$ and $O_P(n^{1-a_S-b_S}) = o_P(1)$ so that $\hat{C}_S = O_P(1)$ and $\hat{\beta}_S$ is inconsistent. More precisely, when $b_S > a_S = 1/2$ the limit in distribution of $\hat{C}_S$ is

$$dlim\ \hat{C}_S = \frac{\lambda_v'\lambda_u}{||\lambda_v||^2}$$

    In addition, note that

$$\begin{aligned}
dlim\ \hat{C}_S &= \frac{\rho}{\sigma_v^2} + \frac{\varepsilon'\lambda_v}{||\lambda_v||^2} \\
dlim\ \hat{C}_S^2 &= \frac{\rho^2}{\sigma_v^4} + \frac{\varepsilon'P_{\lambda_v}\varepsilon}{||\lambda_v||^2} + 2\frac{\rho}{\sigma_v^2}\frac{\varepsilon'\lambda_v}{||\lambda_v||^2}
\end{aligned}$$

$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S) = \frac{\rho}{\sigma_v^2}$$
$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S^2) = \frac{\rho^2}{\sigma_v^4} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)\mathbb{E}(||\lambda_v||^{-2})$$
$$= \frac{\rho^2}{\sigma_v^4} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)(\sigma_v^2(s-2))^{-1}$$

– When $b_S = 1/2$ (S is a locally endogenous IV subset) then $O_P(n^{1-a_S-b_S}) = o_P(1)$ so that $\hat{C}_S = O_P(1)$ and $\hat{\beta}_S$ is inconsistent. More precisely,when $b_S = a_S = 1/2$ the limit in distribution of $\hat{C}_S$ is

$$dlim\ \hat{C}_S = \frac{\lambda_v'(\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_v||^2}$$

In addition, note that

$$dlim\ \hat{C}_S = \frac{\rho}{\sigma_v^2} + \frac{\varepsilon'\lambda_v}{||\lambda_v||^2} + \delta_S'\Sigma_S^{-1/2}\frac{\lambda_v}{||\lambda_v||^2}$$
$$dlim\ \hat{C}_S^2 = \frac{\rho^2}{\sigma_v^4} + \frac{\varepsilon' P_{\lambda_v}\varepsilon}{||\lambda_v||^2} + \delta_S'\Sigma_S^{-1/2}\frac{P_{\lambda_v}}{||\lambda_v||^2}\Sigma_S^{-1/2}\delta_S + 2\frac{\rho}{\sigma_v^2}\frac{\varepsilon'\lambda_v}{||\lambda_v||^2} + 2\delta_S'\Sigma_S^{-1/2}\frac{P_{\lambda_v}}{||\lambda_v||^2}\varepsilon + 2\frac{\rho}{\sigma_v^2}\delta_S'\Sigma_S^{-1/2}\frac{\lambda_v}{||\lambda_v||^2}$$

$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S) = \frac{\rho}{\sigma_v^2} + \delta_S'\Sigma_S^{-1/2}\mathbb{E}\left(\frac{\lambda_v}{||\lambda_v||^2}\right)$$
$$\Rightarrow \mathbb{E}(\ dlim\ \hat{C}_S^2) = \frac{\rho^2}{\sigma_v^4} + 2\frac{\rho}{\sigma_v^2}\delta_S'\Sigma_S^{-1/2}\mathbb{E}\left(\frac{\lambda_v}{||\lambda_v||^2}\right) + \delta_S'\Sigma_S^{-1/2}\mathbb{E}\left(\frac{P_{\lambda_{vS}}}{||\lambda_{vS}||^2}\right)\Sigma_S^{-1/2}\delta_S$$
$$+ \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)\mathbb{E}(||\lambda_{vS}||^{-2})$$
$$= \mathbb{E}\left(\left(\frac{\rho}{\sigma_v^2} + \frac{\lambda_{vS}'}{||\lambda_{vS}||^2}\Sigma_S^{-1/2}\delta_S\right)^2\right) + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_v^2}\right)(\sigma_v^2(s-2))^{-1}$$

– When $b_S < 1/2$ (S is an endogenous IV subset) then $\hat{C}_S$ has no limit and $\hat{\beta}_S$ is inconsistent. But $n^{-1/2}\hat{C}_S = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{ if } b_S = 0 \end{cases}$ and more precisely if $b_S = 0$ then $dlim\ n^{-1/2}\hat{C}_S = \frac{\lambda_v\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}$.

□

Lemma 3.2

Under Assumption A, for any $S \in \mathcal{S}$ *then* $\hat{R}_{EXO,o}(S) = \left(O_P(n^{-1/2}) + O_P(n^{-b_S})\right)^2$ *and*

- *If* $b_S > 0$ *then plim* $\hat{R}_{EXO,o}(S) = \lim R_{EXO}(S) = 0$
- *If* $b_S = 0$ *then*
  - *If* $a_S < 1/2$ *then plim* $\hat{R}_{EXO,o}(S) = \lim R_{EXO}(S) = ||\Sigma_S^{-1/2}\delta_S - \frac{\delta_S'\kappa_S}{\kappa_S'\Sigma_S\kappa_S}\Sigma_S^{-1/2}\delta_S||^2$
  - *If* $a_S = 1/2$ *then dlim* $\hat{R}_{EXO,o}(S) = ||\Sigma_S^{-1/2}\delta_S - \frac{\lambda_{vS}'\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}\lambda_{vS}^*||^2$
  - *If* $a_S > 1/2$ *then dlim* $\hat{R}_{EXO,o}(S) = ||\Sigma_S^{-1/2}\delta_S - \frac{\lambda_v'\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}\lambda_v^*||^2$

Proof. With a slight abuse of the O_P notation

$$\begin{aligned}
\hat{R}_{EXO,o}(S) &= \frac{1}{n}(u_S^* - x^*\hat{C}_S)'z_S^*\Sigma_S^{-1}\frac{1}{n}z_S^{*'}(u_S^* - x^*\hat{C}_S) \\
&= \frac{1}{n}(z_{\bar{E}}^*\alpha + u^* - \hat{C}_S(z_S^*\pi_S + v^*))'z_S^*\Sigma_S^{-1}\frac{1}{n}z_S^{*'}(z_{\bar{E}}^*\alpha + u^* - \hat{C}_S(z_S^*\pi_S + v^*)) \\
&= \left(O_P(n^{-1/2}) + O_P(n^{-b_S}) - \hat{C}_S\left(O_P(n^{-a_S}) + O_P(n^{-1/2})\right)\right)^2
\end{aligned}$$

Then going case by case, using the LLN, the CMT and results from the proof of Lemma 3.1, the limit of $\hat{R}_{EXO,o}(S)$ can be found for any S

- When $a_S < 1/2$ then $\hat{C}_S = O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1/2 - b_S}) + o_P(1)$ and

$$\begin{aligned}
\hat{R}_{EXO,o}(S) = \Big(&O_P(n^{-1/2}) + O_P(n^{-b_S}) - O_P(n^{-b_S}) - 2O_P(n^{a_S - b_S - 1/2}) - O_P(n^{2a_S - 1 - b_S}) \\
&- o_P(1)\left(O_P(n^{-a_S}) + O_P(n^{-1/2})\right)\Big)^2 \\
= &\left(O_P(n^{-1/2}) + O_P(n^{-b_S})\right)^2
\end{aligned}$$

  where the last equality holds by the O_P rules and because $a_S - b_S - 1/2 < -b_S$ and $2a_S - b_S - 1 < -b_S$. Then case by case

  - If $b_S > a_S$ then $\hat{C}_S = o_P(1)$ so that $\hat{R}_{EXO,o}(S) = o_P(1)$

– If $b_S = a_S$ then $\hat{C}_S = O_P(1)$ so that $\hat{R}_{EXO,o}(S) = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{ if } b_S = 0 \end{cases}$. More precisely when $b_S = a_S = 0$ then $plim\ \hat{C}_S = \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$ so that

$$\begin{aligned} plim\ \hat{R}_{EXO,o}(S) &= \left(\Sigma_S^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S \right)' \left(\Sigma^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S \right) \\ &= ||\Sigma_S^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S||^2 \end{aligned}$$

– If $b_S < a_S$ then $\hat{C}_S$ has no limit however because $\hat{C}_S = O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1/2 - b_S}) + o_P(1)$ then

$$\begin{aligned} \hat{R}_{EXO,o}(S) &= \Big(O_P(n^{-1/2}) + O_P(n^{-b_S}) - O_P(n^{-b_S}) - 2O_P(n^{a_S - b_S - 1/2}) - O_P(n^{2a_S - 1 - b_S}) \\ &\quad - o_P(1) \left(O_P(n^{-a_S}) + O_P(n^{-1/2}) \right) \Big)^2 \\ &= \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{ if } b_S = 0 \end{cases} \end{aligned}$$

More precisely when $b_S = 0 < a_S$ then $plim\ n^{-a_S} \hat{C}_S = \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$ so that once again

$$\begin{aligned} plim\ \hat{R}_{EXO,o}(S) &= \left(\Sigma_S^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S \right)' \left(\Sigma^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S \right) \\ &= ||\Sigma_S^{-1/2} \delta_S - \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \Sigma_S^{1/2} \kappa_S||^2 \end{aligned}$$

- When $a_S = 1/2$ then $\hat{C}_S = O_P(n^{1/2 - b_S}) + O_P(1) + o_P(1)$ and

$$\begin{aligned} \hat{R}_{EXO,o}(S) &= \left(O_P(n^{-1/2}) + O_P(n^{-b_S}) - O_P(n^{-b_S}) - (O_P(1) + o_P(1)) O_P(n^{-1/2}) \right)^2 \\ &= \left(O_P(n^{-1/2}) + O_P(n^{-b_S}) \right)^2 \end{aligned}$$

where the last equality holds by the O_P rules. Then case by case

– If $b_S \geqslant 1/2$ then $\hat{C}_S = O_P(1)$ so that $\hat{R}_{EXO,o}(S) = o_P(1)$

– If $b_S < 1/2$ then $\hat{C}_S$ has no limit however

$$\hat{R}_{EXO,o}(S) = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{ if } b_S = 0 \end{cases}$$

More precisely when $b_S = 0$ then $dlim\ n^{-1/2}\hat{C}_S = \frac{\lambda'_{vS}\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}$ so that

$$dlim\ \hat{R}_{EXO,o}(S) = \left(\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_{vS}\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}\lambda^*_{vS}\right)' \left(\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_{vS}\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}\lambda^*_{vS}\right)$$
$$= ||\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_{vS}\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}\lambda^*_{vS}||^2$$

- When $a_S > 1/2$ then $\hat{C}_S = O_P(n^{1-a_S-b_S}) + O_P(n^{1/2-b_S}) + O_P(1) + O_P(n^{1/2-a_S})$ and

$$\hat{R}_{EXO,o}(S) = \Big(O_P(n^{-1/2}) + O_P(n^{-b_S}) - O_P(n^{1-2a_S-b_S}) - 2O_P(n^{1/2-a_S-b_S}) - O_P(n^{-b_S})$$
$$- (O_P(n^{-a_S}) + O_P(n^{-1/2}))(O_P(1) + O_P(n^{1/2-a_S}))\Big)^2$$
$$= \Big(O_P(n^{-1/2}) + O_P(n^{-b_S})\Big)^2$$

  where the last equality holds by the O_P rules and because $1 - 2a_S - b_S < -b_S$ and $1/2 - a_S - b_S < -b_S$. Then case by case

  - If $b_S \geqslant a_S$ then $\hat{C}_S = O_P(1)$ so that $\hat{R}_{EXO,o}(S) = o_P(1)$
  - If $b_S < a_S$ then $\hat{C}_S$ has no limit however

$$\hat{R}_{EXO,o}(S) = \begin{cases} o_P(1) \text{ if } b_S > 0 \\ O_P(1) \text{ if } b_S = 0 \end{cases}$$

  More precisely when $b_S = 0$ then $dlim\ n^{-a_S}\hat{C}_S = \frac{\lambda'_v\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}$ so that

$$dlim\ \hat{R}_{EXO,o}(S) = \left(\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_v\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}\lambda^*_v\right)' \left(\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_v\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}\lambda^*_v\right)$$
$$= ||\Sigma_S^{-1/2}\delta_S - \frac{\lambda'_v\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}\lambda^*_v||^2$$

Then, from D.3 it can be shown that for any $S : a_S < 1/2$, $\mathbb{E}(\ dlim\ \hat{R}_{EXO,o}(S)) = \ lim\ R_{EXO}(S)$. Indeed,

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z^*_S z^{*'}_{\overline{E}})\alpha - \Sigma_S^{1/2}\pi_S\hat{C}_S||^2\right)$$

The statement of the lemma is obtained by combining the results derived above. □

Lemma 3.3

Under Assumption A, for any $S \in \mathcal{S}$ then $n\hat{R}_{EXO,o}(S) = \left(O_P(n^{1/2-b_S}) + O_P(1)\right)^2$ *and*

- *If* $b_S < 1/2$ *then* $n\hat{R}_{EXO,o}(S)$ *and* $\tilde{R}_{EXO}(S)$ *have no limit and explode*
- *If* $b_S = 1/2$ *then* $\mathbb{E}(\ dlim\ n\hat{R}_{EXO,o}(S)) = \ lim\ \tilde{R}_{EXO}(S)$ *and*
  - *If* $a_S < 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\Sigma_S^{1/2}\kappa_S}{\kappa_S'\Sigma_S\kappa_S}\Sigma_S^{1/2}\kappa_S||^2$
  - *If* $a_S = 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\lambda_{vS}}{||\lambda_{vS}||^2}\lambda_{vS}^*||^2$
  - *If* $a_S > 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\lambda_v}{||\lambda_v||^2}\lambda_v^*||^2$
- *If* $b_S > 1/2$ *then* $\mathbb{E}(\ dlim\ n\hat{R}_{EXO,o}(S)) = \ lim\ \tilde{R}_{EXO}(S)$ *and*
  - *If* $a_S < 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\Sigma_S^{1/2}\kappa_S}{\kappa_S'\Sigma_S\kappa_S}\Sigma_S^{1/2}\kappa_S||^2$
  - *If* $a_S = 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\lambda_{vS}}{||\lambda_{vS}||^2}\lambda_{vS}^*||^2$
  - *If* $a_S > 1/2$ *then* $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\lambda_v}{||\lambda_v||^2}\lambda_v^*||^2$

Proof. With a slight abuse of the O_P notation

$$\begin{aligned}
n\hat{R}_{EXO,o}(S) &= \frac{1}{\sqrt{n}}(u_S - x\hat{C}_S)'z_S\Sigma_S^{-1}\frac{1}{\sqrt{n}}z_S'(u_S - x\hat{C}_S) \\
&= \frac{1}{\sqrt{n}}(z_{\overline{E}}\alpha + u - \hat{C}_S(z_S\pi_S + v))'z_S\Sigma_S^{-1}\frac{1}{\sqrt{n}}z_S'(z_{\overline{E}}\alpha + u - \hat{C}_S(z_S\pi_S + v)) \\
&= \left(O_P(n^{1/2-b_S}) + O_P(1) - \hat{C}_S\left(O_P(n^{1/2-a_S}) + O_P(1)\right)\right)^2
\end{aligned}$$

Then case by case, using the LLN, the CMT and results from the proofs of Lemma 3.1 and Lemma 3.2 the limit of $\hat{R}_{EXO,o}(S)$ can be found for any S

- When $a_S < 1/2$
  - If $b_S > 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\Sigma_S^{1/2}\kappa_S}{\kappa_S'\Sigma_S\kappa_S}\Sigma_S^{1/2}\kappa_S||^2$
  - If $b_S = 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\Sigma_S^{1/2}\kappa_S}{\kappa_S'\Sigma_S\kappa_S}\Sigma_S^{1/2}\kappa_S||^2$
  - If $b_S < 1/2$ then $n\hat{R}_{EXO,o}(S) = O_P(n^{1/2-b_S})$ has no limit

- When $a_S = 1/2$
  - If $b_S > 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\lambda_{vS}}{||\lambda_{vS}||^2}\lambda_{vS}^*||^2$
  - If $b_S = 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\lambda_{vS}}{||\lambda_{vS}||^2}\lambda_{vS}^*||^2$
  - If $b_S < 1/2$ then $n\hat{R}_{EXO,o}(S) = O_P(n^{1/2-b_S})$ has no limit
- When $a_S > 1/2$
  - If $b_S > 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* - \frac{\lambda_u'\lambda_v}{||\lambda_v||^2}\lambda_v^*||^2$
  - If $b_S = 1/2$ then $dlim\ n\hat{R}_{EXO,o}(S) = ||\lambda_u^* + \Sigma_S^{-1/2}\delta_S - \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)'\lambda_v}{||\lambda_v||^2}\lambda_v^*||^2$
  - If $b_S < 1/2$ then $n\hat{R}_{EXO,o}(S) = O_P(n^{1/2-b_S})$ has no limit

Then, from D.3 it can be again shown that for any $S : b_S \geqslant 1/2$, $\mathbb{E}(\ dlim\ n\hat{R}_{EXO,o}(S)) = lim\ \tilde{R}_{EXO}(S)$ and that for any $S : b_S < 1/2$, $\tilde{R}_{EXO}(S)$ also explodes. The statement of the lemma is obtained by combining the results derived above. □

Lemma 3.4
Under Assumption A, for any $S \in \mathcal{S}$ then $\hat{R}_{PMSE,o}(S) = O_P(1) + O_P(n^{-2b_S})$

- *If $b_S > 0$ then $plim\ \hat{R}_{PMSE,o}(S) = \ lim\ R_{PMSE}(S) = \mathbb{E}((v^*\beta + u^*)^2)$*
- *If $b_S = 0$ then $\mathbb{E}(\ dlim\ \hat{R}_{PMSE,o}(S)) = \ lim\ R_{PMSE}(S)$ and*
  - *If $a_S < 1/2$ then $plim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2) + \alpha'\Sigma_{\overline{E}}\alpha - \delta_S'\Sigma_S^{-1/2}P_{\Sigma_S^{1/2}\kappa_S}\Sigma_S^{-1/2}\delta_S$*
  - *If $a_S = 1/2$ then $dlim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2) + \alpha'\Sigma_{\overline{E}}\alpha + \delta_S'\Sigma_S^{-1/2}P_{\lambda_{vS}}\Sigma_S^{-1/2}\delta_S - 2\delta_S'\kappa_S\frac{\lambda_{vS}'\Sigma_S^{-1/2}\delta_S}{||\lambda_{vS}||^2}$*
  - *If $a_S > 1/2$ then $dlim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2) + \alpha'\Sigma_{\overline{E}}\alpha + \delta_S'\Sigma_S^{-1/2}P_{\lambda_v}\Sigma_S^{-1/2}\delta_S - 2\delta_S'\kappa_S\frac{\lambda_v'\Sigma_S^{-1/2}\delta_S}{||\lambda_v||^2}$*

Proof. With a slight abuse of the O_P notation

$$\begin{aligned}\hat{R}_{PMSE,o}(S) &= \frac{1}{n}||v^*\beta + u^* + z_{\overline{E}}^*\alpha - z_S^*\pi_S\hat{C}_S||^2 \\ &= \frac{1}{n}||v^*\beta + u^*||^2 + \frac{1}{n}||z_{\overline{E}}^*\alpha - z_S^*\pi_S\hat{C}_S||^2 + \frac{2}{n}(v^*\beta + u^*)'(z_{\overline{E}}^*\alpha - z_S^*\pi_S\hat{C}_S) \\ &= O_P(1) + O_P(n^{-2b_S}) + O_P(n^{-2a_S})\hat{C}_S^2 - 2O_P(n^{-a_S-b_S})\hat{C}_S + 2O_P(n^{-1/2-b_S}) - 2O_P(n^{-1/2-a_S})\hat{C}_S\end{aligned}$$

- When $a_S < 1/2$ then $\hat{C}_S = O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1/2 - b_S}) + o_P(1)$ and

$$\hat{R}_{PMSE,o}(S) = O_P(1) + O_P(n^{-2b_S}) + O_P(n^{-a_S - b_S}) - 2O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > a_S$ then $\hat{C}_S = o_P(1)$ so that $plim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2)$
  - If $b_S = a_S$ then $plim\ \hat{C}_S = \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$ so that

$$plim\ \hat{R}_{PMSE,o}(S) = \begin{cases} \mathbb{E}((v^*\beta + u^*)^2) \text{ if } b_S = a_S > 0 \\ \mathbb{E}((v^*\beta + u^*)^2) + \alpha' \Sigma_{\overline{E}} \alpha - \delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2} \kappa_S} \Sigma_S^{-1/2} \delta_S \text{ if } b_S = a_S = 0 \end{cases}$$

  - If $b_S < a_S$ then $plim\ n^{-a_S} \hat{C}_S = \begin{cases} 0 \text{ if } a_S > b_S > 0 \\ \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} \text{ if } a_S > b_S = 0 \end{cases}$ so that

$$plim\ \hat{R}_{PMSE,o}(S) = \begin{cases} \mathbb{E}((v^*\beta + u^*)^2) \text{ if } a_S > b_S > 0 \\ \mathbb{E}((v^*\beta + u^*)^2) + \alpha' \Sigma_{\overline{E}} \alpha - \delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2} \kappa_S} \Sigma_S^{-1/2} \delta_S \text{ if } a_S > b_S = 0 \end{cases}$$

- When $a_S = 1/2$ then $\hat{C}_S = O_P(n^{1/2 - b_S}) + O_P(1) + o_P(1)$ and

$$\hat{R}_{PMSE,o}(S) = O_P(1) + O_P(n^{-2b_S}) + O_P(n^{-2b_S}) - 2O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > 1/2$ then $\hat{C}_S = O_P(1)$ so that $plim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2)$
  - If $b_S = 0$ then $dlim\ n^{-1/2} \hat{C}_S = \frac{\lambda_{vS}' \Sigma_S^{-1/2} \delta_S}{||\lambda_{vS}||^2}$ so that

$$dlim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2) + \alpha' \Sigma_{\overline{E}} \alpha + \delta_S' \Sigma_S^{-1/2} P_{\lambda_{vS}} \Sigma_S^{-1/2} \delta_S - 2\delta_S' \kappa_S \frac{\lambda_{vS}' \Sigma_S^{-1/2} \delta_S}{||\lambda_{vS}||^2}$$

- When $a_S > 1/2$ then $\hat{C}_S = O_P(n^{1 - a_S - b_S}) + O_P(n^{1/2 - b_S}) + o_P(1)$ and

$$\hat{R}_{PMSE,o}(S) = O_P(1) + O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > 0$ then $\hat{C}_S = O_P(1)$ so that $plim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2)$
  - If $b_S = 0$ then $dlim\ n^{-1/2} \hat{C}_S = \frac{\lambda_v' \Sigma_S^{-1/2} \delta_S}{||\lambda_v||^2}$ so that

$$dlim\ \hat{R}_{PMSE,o}(S) = \mathbb{E}((v^*\beta + u^*)^2) + \alpha' \Sigma_{\overline{E}} \alpha + \delta_S' \Sigma_S^{-1/2} P_{\lambda_v} \Sigma_S^{-1/2} \delta_S - 2\delta_S' \kappa_S \frac{\lambda_v' \Sigma_S^{-1/2} \delta_S}{||\lambda_v||^2}$$

Then, from D.3 it can be again shown that for any S, $\mathbb{E}(\ dlim\ \hat{R}_{PMSE,o}(S)) = \ lim\ R_{PMSE}(S)$. Indeed,

$$R_{PMSE} = \mathbb{E}((v^*\beta + u^*)^2) + \mathbb{E}(||z_{\overline{E}}^{*'} \alpha - z_S^{*'} \pi_S \hat{C}_S||^2)$$

The statement of the lemma is obtained by combining the results derived above. □

Lemma 3.5
Under Assumption A, for any $S \in \mathcal{S}$

- *If $b_S < 1/2$ then*
  - *If $a_S < b_S$ then plim $\hat{R}_{MSE,o}(S) =$ lim $R_{MSE}(S) = \sigma_u^2$*
  - *If $a_S = b_S = 0$ then*

$$\text{plim } \hat{R}_{MSE,o}(S) = \text{ lim } R_{MSE}(S) = \sigma_u^2 + \sigma_v^2 \frac{\delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2}\kappa_S} \Sigma_S^{-1/2} \delta_S}{\kappa_S' \Sigma_S \kappa_S} - 2\rho \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} + \alpha' \Sigma_{\overline{E}} \alpha - \delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2}\kappa_S} \Sigma_S^{-1/2} \delta_S$$

  - *If $0 < a_S = b_S < 1/2$ then*

$$\text{plim } \hat{R}_{MSE,o}(S) = \text{ lim } R_{MSE}(S) = \sigma_u^2 + \sigma_v^2 \frac{\delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2}\kappa_S} \Sigma_S^{-1/2} \delta_S}{\kappa_S' \Sigma_S \kappa_S} - 2\rho \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$$

  - *If $a_S > b_S$ then $\hat{R}_{MSE,o}(S)$ and $R_{MSE}(S)$ have no limit and explode*

  plim $\hat{R}_{MSE,o}(S) =$ lim $R_{MSE}(S)$ have no limit and explode

- *If $b_S = 1/2$ then $\mathbb{E}($ dlim $\hat{R}_{MSE,o}(S)) =$ lim $R_{MSE}(S)$ and*
  - *If $a_S < 1/2$ then plim $\hat{R}_{MSE,o}(S) = \sigma_u^2$*
  - *If $a_S = 1/2$ then dlim $\hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)' P_{\lambda_{vS}} (\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_{vS}||^2} - 2\rho \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)' \lambda_{vS}}{||\lambda_{vS}||^2}$*
  - *If $a_S > 1/2$ then dlim $\hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)' P_{\lambda_v} (\lambda_u + \Sigma_S^{-1/2}\delta_S)}{||\lambda_v||^2} - 2\rho \frac{(\lambda_u + \Sigma_S^{-1/2}\delta_S)' \lambda_v}{||\lambda_v||^2}$*

- *If $b_S > 1/2$ then $\mathbb{E}($ dlim $\hat{R}_{MSE,o}(S)) =$ lim $R_{MSE}(S)$ and*
  - *If $a_S < 1/2$ then dlim $\hat{R}_{MSE,o}(S) = \sigma_u^2$*
  - *If $a_S = 1/2$ then dlim $\hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\lambda_u' P_{\lambda_{vS}} \lambda_u}{||\lambda_{vS}||^2} - 2\rho \frac{\lambda_u' \lambda_{vS}}{||\lambda_{vS}||^2}$*
  - *If $a_S > 1/2$ then dlim $\hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\lambda_u' P_{\lambda_v} \lambda_u}{||\lambda_v||^2} - 2\rho \frac{\lambda_u' \lambda_v}{||\lambda_v||^2}$*

Proof. With a slight abuse of the O_P notation

$$\begin{aligned}
\hat{R}_{MSE,o}(S) &= \frac{1}{n}||u^* - v^* \hat{C}_S + z_{\overline{E}}^* \alpha - z_S^* \pi_S \hat{C}_S||^2 \\
&= \frac{1}{n}||u^* - v^* \hat{C}_S||^2 + \frac{1}{n}||z_{\overline{E}}^* \alpha - z_S^* \pi_S \hat{C}_S||^2 + \frac{2}{n}(u^* - v^* \hat{C}_S)'(z_{\overline{E}}^* \alpha - z_S^* \pi_S \hat{C}_S) \\
&= O_P(1) + O_P(1)\hat{C}_S^2 - 2O_P(1)\hat{C}_S + O_P(n^{-2b_S}) + O_P(n^{-2a_S})\hat{C}_S^2 - 2O_P(n^{-a_S - b_S})\hat{C}_S \\
&\quad + 2O_P(n^{-1/2-b_S}) - 2O_P(n^{-1/2-a_S})\hat{C}_S - 2O_P(n^{-1/2-b_S})\hat{C}_S + 2O_P(n^{-1/2-a_S})\hat{C}_S^2
\end{aligned}$$

- When $a_S < 1/2$ then $\hat{C}_S = O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1/2 - b_S}) + o_P(1)$ and

  $$\hat{R}_{MSE,o}(S) = O_P(1) + O_P(n^{2a_S - 2b_S}) - 2O_P(n^{2a_S - 1/2 - b_S}) + O_P(n^{-2b_S}) + O_P(n^{-a_S - b_S}) - 2O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > a_S$ then $\hat{C}_S = o_P(1)$ so that $plim\ \hat{R}_{MSE,o}(S) = \sigma_u^2$
  - If $b_S = a_S$ then $plim\ \hat{C}_S = \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$ so that if $b_S = a_S > 0$ then

    $$plim\ \hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2} \kappa_S} \Sigma_S^{-1/2} \delta_S}{\kappa_S' \Sigma_S \kappa_S} - 2\rho \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S}$$

    and so that if $b_S = a_S = 0$

    $$plim\ \hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2} \kappa_S} \Sigma_S^{-1/2} \delta_S}{\kappa_S' \Sigma_S \kappa_S} - 2\rho \frac{\kappa_S' \delta_S}{\kappa_S' \Sigma_S \kappa_S} + \alpha' \Sigma_{\overline{E}} \alpha - \delta_S' \Sigma_S^{-1/2} P_{\Sigma_S^{1/2} \kappa_S} \Sigma_S^{-1/2} \delta_S$$

  - If $b_S < a_S$ then $\hat{C}_S$ has no limit so $\hat{R}_{MSE,o}(S)$ has no limit
- When $a_S = 1/2$ then $\hat{C}_S = O_P(n^{1/2 - b_S}) + O_P(1) + o_P(1)$ and

  $$\hat{R}_{MSE,o}(S) = O_P(1) + O_P(n^{1/2 - b_S}) + O_P(1) + O_P(n^{-2b_S}) + O_P(n^{-2b_S}) - 2O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > 1/2$ then $\hat{C}_S = \frac{\lambda_u' \lambda_{vS}}{||\lambda_{vS}||^2}$ so that $dlim\ \hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\lambda_u' P_{\lambda_{vS}} \lambda_u}{||\lambda_{vS}||^2} - 2\rho \frac{\lambda_u' \lambda_{vS}}{||\lambda_{vS}||^2}$
  - If $b_S = 1/2$ then $dlim\ \hat{C}_S = \frac{\lambda_{vS}' (\lambda_u + \Sigma_S^{-1/2} \delta_S)}{||\lambda_{vS}||^2}$ so that

    $$dlim\ \hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{(\lambda_u + \Sigma_S^{-1/2} \delta_S)' P_{\lambda_{vS}} (\lambda_u + \Sigma_S^{-1/2} \delta_S)}{||\lambda_{vS}||^2} - 2\rho \frac{(\lambda_u + \Sigma_S^{-1/2} \delta_S)' \lambda_{vS}}{||\lambda_{vS}||^2}$$

  - If $b_S < 1/2$ then $\hat{C}_S$ has no limit so $\hat{R}_{MSE,o}(S)$ has no limit
- When $a_S > 1/2$ then $\hat{C}_S = O_P(n^{1 - a_S - b_S}) + O_P(n^{1/2 - b_S}) + o_P(1)$ and

  $$\hat{R}_{MSE,o}(S) = O_P(1) + O_P(n^{1 - a_S - b_S}) + O_P(n^{1/2 - b_S}) + O_P(n^{-2b_S}) + o_P(1)$$

  - If $b_S > 1/2$ then $\hat{C}_S = O_P(1)$ so that $dlim\ \hat{R}_{PMSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{\lambda_u' P_{\lambda_v} \lambda_u}{||\lambda_v||^2} - 2\rho \frac{\lambda_u' \lambda_v}{||\lambda_v||^2}$
  - If $b_S = 1/2$ then $dlim\ \hat{C}_S = \frac{\lambda_v' (\lambda_u + \Sigma_S^{-1/2} \delta_S)}{||\lambda_v||^2}$ so that

    $$dlim\ \hat{R}_{MSE,o}(S) = \sigma_u^2 + \sigma_v^2 \frac{(\lambda_u + \Sigma_S^{-1/2} \delta_S)' P_{\lambda_v} (\lambda_u + \Sigma_S^{-1/2} \delta_S)}{||\lambda_v||^2} - 2\rho \frac{(\lambda_u + \Sigma_S^{-1/2} \delta_S)' \lambda_v}{||\lambda_v||^2}$$

  - If $b_S < 1/2$ then $\hat{C}_S$ has no limit so $\hat{R}_{MSE,o}(S)$ has no limit

Then, from D.3 it can be again shown that for any $S : b_S \geqslant 1/2$ or any $S : a_S \leqslant b_S < 1/2$,, $\mathbb{E}(\ dlim\ \hat{R}_{MSE,o}(S)) = \ lim\ R_{PMSE}(S)$. The statement of the lemma is obtained by combining the results derived above. □

Lemma 3.6
Under Assumptions A and D

- *For $k = \{r; c; an\}$ if $S \in \mathcal{S}_k$ then for any $S' \notin \mathcal{S}_k$*

$$lim\ \frac{\overline{R}_{EXO}(S') - \overline{R}_{EXO}(S)}{\overline{R}_{EXO}(S)} > 0$$

- *For $k = \{r; c; an\}$ if $S \in \mathcal{S}_k$ then for any $S' \notin \mathcal{S}_k$*

$$lim\ \frac{R_{PMSE}(S') - R_{PMSE}(S)}{R_{PMSE}(S) - (\sigma_u^2 + \sigma_v^2\beta^2 + 2\rho\beta)} > 0$$

- *If $S \in \mathcal{S}_c$ then for any $S' \notin \mathcal{S}_c$*

$$lim\ \frac{R_{MSE}(S') - R_{MSE}(S)}{R_{MSE}(S) - \sigma_u^2} > 0$$

Proof. The proof is in two parts, first I prove that endogenous sets have higher risks, second I prove that strong sets have higher risks.

Let $S \in \mathcal{S}_r$ so that $b_S > 1/2$. Then from lemmas 3.3 and 3.5 for any $S' : b_{S'} < 1/2$ then $\tilde{R}_{EXO}(S')$ and $R_{MSE}(S')$ explode thus $lim\ \tilde{R}_{EXO}(S) - \tilde{R}_{EXO}(S') = -\infty$ and $lim\ R_{MSE}(S) - R_{MSE}(S') = -\infty$. Furthermore, for any $S' : b_{S'} = 1/2$ then $\tilde{R}_{EXO}(S')$ is equal to the mean of a non-central chi-square distribution with $s' + 1$ degrees of freedom, and it is strictly superior to $\tilde{R}_{EXO}(S)$ which is equal to the mean of a non-central chi-square distribution with $s' + 1$ degrees of freedom, thus $lim\ \tilde{R}_{EXO}(S) - \tilde{R}_{EXO}(S') < 0$. On the other hand, for any S', $R_{EXO}(S)' = O(n^{-2b_{S'}})$, $R_{PMSE}(S') = \mathbb{E}((v^*\beta + u^*)^2) + O(n^{-2b_{S'}})$, and $R_{MSE}(S') = \sigma_u^2 + \sigma_v^2 O(n^{1-2b_{S'}}) - 2\rho O(n^{1/2-b_{S'}})$ thus for any $S' : b_S = 1/2$, for $k \in \{EXO; PMSE; MSE\}$, $lim\ R_k(S) - R_k(S') < 0$. This proves the first bullet point of the lemma. □

C.2 Technical Lemmas: convergence of risk estimators

Lemma 3.7

Consider B samples $((X_{i,b})_{i=1}^{n_b})_{b=1}^{B}$ coming from the original sample $(X_i)_{i=1}^{n}$ of random variables $(X_i)_{i=1}^{n}$ and a resampled statistic $\hat{S}_b = f((X_{i,b})_{i=1}^{n_b})$ such that $Var(\hat{S}_b) < +\infty$, $((X_{i,b})_{i=1}^{n_b})_{b=1}^{B}$ is identically distributed across b, $n_b \underset{n\to+\infty}{\to} +\infty$, and $B \underset{n\to+\infty}{\to} +\infty$. Then

$$plim \left| \frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b | (X_i)_{i=1}^{n}) \right| = 0$$

Furthermore, if the B samples are independent or if for any b, $\sum_{b'=1}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_b} Var(\hat{S}_b) c^{n_b - n^*}$ for some $c \in (0;1)$ then

$$plim \left| \frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b) \right| = 0$$

Proof. By Chebyshev's inequality and using the fact that conditionally on $(X_i)_{i=1}^{n}$, $((X_{i,b})_{i=1}^{n_b})_{b=1}^{B}$ and therefore $(\hat{S}_b)_{b=1}^{B}$ is iid across b, for any $e > 0$

$$\begin{aligned}
\mathbb{P}(|\frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b | (X_i)_{i=1}^{n})| > e | (X_i)_{i=1}^{n}) &\leqslant \frac{\frac{1}{B^2} Var(\sum_{b=1}^{B} \hat{S}_b | (X_i)_{i=1}^{n})}{e^2} \\
&= \frac{\frac{1}{B} Var(\hat{S}_b | (X_i)_{i=1}^{n})}{e^2} \\
\Rightarrow \mathbb{P}(|\frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b | (X_i)_{i=1}^{n})| > e) &\leqslant \frac{\frac{1}{B} \mathbb{E}(Var(\hat{S}_b | (X_i)_{i=1}^{n}))}{e^2} \to 0
\end{aligned}$$

because $\mathbb{E}(Var(\hat{S}_b | (X_i)_{i=1}^{n})) \leqslant Var(\hat{S}_b) < +\infty$. On the other hand note that

$$\begin{aligned}
&\sum_{b'=1}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_b} Var(\hat{S}_b) c^{n_b - n^*} = Var(\hat{S}_b) c^{n_b} \left(1 + \frac{1 - c^{-n_b - 1}}{1 - c^{-1}}\right) = Var(\hat{S}_b) \left(c^{n_b} + \frac{c^{n_b} - c^{-1}}{1 - c^{-1}}\right) \\
&\Rightarrow \frac{1}{B^2} \sum_{b,b'}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant Var(\hat{S}_b) \left(\frac{c^{n_b}}{B} + \frac{\frac{c^{n_b}}{B} - \frac{1}{Bc}}{1 - \frac{1}{c}}\right) \to 0
\end{aligned}$$

Consequently, if the B samples are independent then $Cov(\hat{S}_b, \hat{S}_{b'}) = 0$ so that

$$\mathbb{P}\left(\left|\frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b)\right| > e\right) \leqslant \frac{\frac{1}{B} Var(\hat{S}_b)}{e^2} \to 0$$

And if $\sum_{b'=1}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_b} Var(\hat{S}_b) c^{n_b - n^*}$ for any (b, b') then

$$\mathbb{P}(|\frac{1}{B} \sum_{b=1}^{B} \hat{S}_b - \mathbb{E}(\hat{S}_b)| > e) \leqslant \frac{\frac{1}{B^2} \sum_{b,b'} Cov(\hat{S}_b, \hat{S}_{b'})}{e^2} \to 0$$

□

Lemma 3.8
Under assumptions A and B, for $k \in \{EXO; PMSE; MSE\}$, for any S, $\hat{R}_k(S)$ is uniformly integrable and has finite second moment. Under the same conditions, for any S such that $b_S \geqslant 1/2$, $n_b \hat{R}_{EXO}(S)$ is uniformly integrable and has finite second moment.

Proof. The proof is straightforward. First, from the previous lemmas 3.2, 3.3, 3.4, 3.5, all the criteria are bounded in probability for any S and $n_b \hat{R}_{EXO}(S)$ is bounded in probability for any $S : b_S \geqslant 1/2$. Second, using assumptions A and B (more specifically that the data has finite moments of order 4, and the fact that the denominator of 2SLS is bounded away from 0) and Cauchy-Schwarz inequality, it is simple to prove that for $k \in \{EXO; PMSE; MSE\}$, for any S, $\mathbb{E}(\hat{R}_k(S)^2) < +\infty$ and similarly that for any $S : b_S \geqslant 1/2$, $\mathbb{E}(n_b^2 \hat{R}_{EXO}(S)^2) < +\infty$. Third, as a direct consequence, uniform integrability holds for any S (such that $b_S \geqslant 1/2$ in case of $n_b \hat{R}_{EXO}(S)$)

$$\mathbb{E}(\hat{R}_k(S) 1\{(\hat{R}_k(S) > e\}) \leqslant \mathbb{E}(\hat{R}_k^2(S)) \mathbb{P}(\hat{R}_k(S) > e) \underset{e \to +\infty}{\longrightarrow} 0$$
$$\mathbb{E}(n_b \hat{R}_{EXO}(S) 1\{(n_b \hat{R}_{EXO}(S) > e\}) \leqslant \mathbb{E}(n_b^2 \hat{R}_{EXO}^2(S)) \mathbb{P}(n_b \hat{R}_{EXO}(S) > e) \underset{e \to +\infty}{\longrightarrow} 0$$

□

Lemma 3.9
Under assumptions A, B, and C, for $k \in \{EXO; PMSE; MSE\}$ and for any S

$$plim\ |\hat{R}_k(S) - R_k(S)| = 0$$

moreover under the same conditions, for any $S : b_S \geqslant 1/2$

$$plim\ |n_b \hat{R}_{EXO}(S) - \tilde{R}_{EXO}(S))| = 0$$

Proof. First, note that for $k \in \{EXO; PMSE; MSE\}$ and for any $e > 0$

$$\begin{aligned}\mathbb{P}(|\hat{R}_k(S) - R_k(S)| > e) \leqslant\ & \mathbb{P}(|\hat{R}_k(S) - \mathbb{E}(\hat{R}_k(S))| > e/2) \\ & + \mathbb{P}(|\mathbb{E}(\hat{R}_k(S)) - R_k(S)| > e/2)\end{aligned}$$

because $\mathbb{P}(A + B > e) \leqslant \mathbb{P}(A > e/2) + \mathbb{P}(B > e/2)$ for any random variables (A, B).

Next, the second term on the right converges to zero because

$$lim\ |\mathbb{E}(\hat{R}_k(S)) - R_k(S)| \leqslant\ lim\ |\mathbb{E}(\hat{R}_k(S)) - \mathbb{E}(\hat{R}_{k,o}(S))| +\ lim\ |\mathbb{E}(\hat{R}_{k,o}(S)) - R_k(S)| = 0$$

Indeed by Lemma 3.2, Lemma 3.4, and Lemma 3.5, $lim\ |\mathbb{E}(\hat{R}_{k,o}(S)) - R_k(S)| = 0$, and due to the fact that the resampled data is identically distributed $\mathbb{E}(\hat{R}_k(S)) = \frac{1}{B}\sum_b \mathbb{E}(\hat{R}_{k,b}) = \mathbb{E}(\hat{R}_{k,b}) = \mathbb{E}(\hat{R}_{k,o})$ so that $lim\ |\mathbb{E}(\hat{R}_k(S)) - \mathbb{E}(\hat{R}_{k,o}(S))| = 0$. Then (almost) sure convergence implies convergence in probability.

Next, by Lemma 3.7 whose assumptions are satisfied

$$\mathbb{P}(|\hat{R}_k(S) - \mathbb{E}(\hat{R}_k(S))| > e/2) = \mathbb{P}\left(\left|\frac{1}{B}\sum_{b=1}^{B}\hat{R}_{k,b}(S) - \mathbb{E}(\hat{R}_{k,b}(S))\right| > e/2\right) \to 0$$

Therefore, for $k \in \{PMSE; MSE\}$, for any S

$$plim\ |\hat{R}_k(S) - R_k(S)| = 0$$

With a similar argument, using Lemma 3.3 on the limit of $\tilde{R}_{EXO}(S))$ and Lemma 3.7 on the convergence of risk estimators to their mean, for any S

$$plim\ |n_b\hat{R}_{EXO}(S) - \tilde{R}_{EXO}(S))| = 0$$

□

C.3 Proof of Theorem 4.1

The proof is in 2 steps. First, I prove that for $k \in \{EXO; PMSE; MSE\}$ for any $S \in \mathcal{S}$, $plim\ \frac{|\hat{R}_k(S) - R_k(S)|}{R_k(S)} = 0$. Second, I prove that this implies $plim\ \frac{\hat{R}_k(\hat{S}_{\hat{R}_k})}{\min\limits_{S\in\mathcal{S}} R_k(S)} = 1$.

- If $k \in \{PMSE; MSE\}$ notice that for any $\tilde{\beta} \in \mathbb{R}$

  $$\mathbb{E}((u^* - v^*\tilde{\beta})^2) = 0 \Leftrightarrow \sigma_u^2 + \sigma_v^2\tilde{\beta}^2 - 2\rho\tilde{\beta}$$

  But discriminant $4(\rho^2 - \sigma_u^2\sigma_v^2)$ is strictly negative thus $\forall\tilde{\beta} \in \mathbb{R}\ \mathbb{E}((u^* - v^*\tilde{\beta})^2) > 0$. Hence for $k \in \{PMSE; MSE\}$, looking at the decomposition of $R_k(S)$ from section 4.1, there exists $c > 0$ such that $\forall S\ \mathbb{P}(R_k(S) \leqslant c) = 0$. Consequently, for any $S \in \mathcal{S}$, for any $e > 0$

  $$\mathbb{P}\left(\left|\frac{\hat{R}_k(S) - R_k(S)}{R_k(S)}\right| > e\right) \leqslant \mathbb{P}\left(|\hat{R}_k(S) - R_k(S)| > ec\right)$$

which converges to zero by Lemma 3.9.

If $k = EXO$, then

$$\begin{aligned}\tilde{R}_{EXO}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*'}\Sigma_S^{-1}z_S^*(y^* - x^*\hat{\beta}_S)\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}_n\left(||\Sigma_S^{-1/2}z_S^*(u^* - v^*\hat{C}_S)||^2\right)\right) + \pi_S'\Sigma_S\pi_S\mathbb{E}_n\left(\hat{C}_S^2\right)\end{aligned}$$

which is strictly positive because the first term is strictly positive. Similarly, when $b_S = 0$ then

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z_S^*z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2}\pi_S\hat{C}_S||^2\right)$$

is also strictly positive.

Consequently, for any S there exists some $c > 0$ such that

$$\mathbb{P}\left(\left|\frac{\hat{R}_{EXO}(S) - \overline{R}_{EXO}(S)}{\overline{R}_{EXO}(S)}\right| > e\right) \leqslant \mathbb{P}\left(|\hat{R}_{EXO}(S) - \overline{R}_{EXO}(S)| > ec\right)$$

which converges to zero from Lemma 3.9.

- If for any S and for $k \in \{PMSE; MSE\}$, $plim\ \frac{\hat{R}_k(S) - R_k(S)}{R_k(S)} = 0$, then $plim \max_{S\in\mathcal{S}} \frac{\hat{R}_k(S) - R_k(S)}{R_k(S)} = 0$ because $\mathcal{S}$ is finite. Then denote by $S^* = Argmin\ R_k(S)$ and notice that

$$\hat{R}_k(S^*) \geqslant R_k(S^*), \qquad \hat{R}_k(S^*) \geqslant \hat{R}_k(\hat{S}_{\hat{R}_k}), \qquad (R_k(\hat{S}_{\hat{R}_k}))^{-1} \leqslant (\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*))^{-1}$$

for some $\alpha \in (0;1)$. Therefore,

$$\begin{aligned}\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} &\leqslant \frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \leqslant \frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*) + \hat{R}_k(S^*) - \hat{R}_k(\hat{S}_{\hat{R}_k})}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \\ &\leqslant \frac{|R_k(\hat{S}_{\hat{R}_k}) - \hat{R}_k(\hat{S}_{\hat{R}_k})| + |\hat{R}_k(S^*) - R_k(S^*)|}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \\ &\leqslant \frac{2}{\alpha}\max_{S\in\mathcal{S}}\frac{|\hat{R}_k(S) - R_k(S)|}{R_k(S)} \xrightarrow{\mathbb{P}} 0\end{aligned}$$

On the other hand because $R_k(\hat{S}_{\hat{R}_k}) \geqslant R_k(S^*)$, $\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} \geqslant 0$ thus

$$\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} = 1 - \frac{R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} \xrightarrow{\mathbb{P}} 0$$

By the CMT this implies $\frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} \xrightarrow{\mathbb{P}} 1$ thus again by the CMT

$$\frac{\hat{R}_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} = \frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} + \frac{\hat{R}_k(\hat{S}_{\hat{R}_k}) - R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} = \frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} + o_P(1) \xrightarrow{\mathbb{P}} 1$$

With the exact same arguments, it can be proven that $plim\ \frac{\hat{R}_{EXO}(\hat{S}_{\hat{R}_{EXO}})}{\min\limits_{S\in\mathcal{S}} \overline{R}_{EXO}(S)} = 1$ because $\overline{R}_{EXO}(S)$ is equal to $\frac{1}{n_b}\tilde{R}_{EXO}(S)$ when $b_S \geqslant 1/2$ and is equal to $R_{EXO}(S)$ otherwise.

C.4 Proof of Theorem 4.2

For $k \in \{EXO; PMSE; MSE\}$ and using $\overline{R}_{EXO}$ as the target risk in case $k = EXO$, from Lemma 3.6 on the rankings between IV sets based on different expected risks, if there exists some $S \in \mathcal{S}_c$ then for any $S \in \mathcal{S}_c$, for any $S' \notin \mathcal{S}_c$ the inequality $lim\ R_k(S) < lim\ R_k(S')$ holds so that $lim\ R_k(S) = lim \min\limits_{S\in\mathcal{S}} R_k(S)$. At the same time, from Lemma 3.9 and Theorem 4.1 on the convergence of risk estimators $plim \min\limits_{S} \hat{R}_k(S) = lim \min\limits_{S} R_k(S)$. As a consequence, $plim\ \hat{S}_{\hat{R}_k} = plim\ Argmin\limits_{S\in\mathcal{S}} \hat{R}_k(S) = lim\ Argmin\limits_{S\in\mathcal{S}} R_k(S)$ and at the limit $Argmin\limits_{S\in\mathcal{S}} R_k(S)$ belongs to $\mathcal{S}_c$ (surely), therefore $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_c$ with probability one at the limit. The exact same reasoning can be applied to prove that if there exists some $S \in \mathcal{S}_{an}$ then $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_{an}$ with probability one at the limit, and that if there exists some $S \in \mathcal{S}_r$ then $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_r$ with probability one at the limit.

D Additional theoretical results

In this section, are formal results of statements made in the paper. First, are formal characterizations of the target parameter sets from section 2.3. Second are formal proofs that the "true models" from section 3.1 are well defined and well characterized by conditions on a_S and b_S. Third is the step-by-step decomposition of the risks from section 4.1. Fourth, I provide different assumptions for the consistency of the risk estimators using k-class estimators instead of 2SLS. Refer to appendix C for notations and conventions.

D.1 IV estimation target sets

In this subsection, I characterize the target parameter set in case the objective behind IV estimation is minimizing the weighted sum of exclusion restrictions or minimizing the mean squared error after projection of the endogenous variable on the IVs from section 2.3.

First, let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$. It turns out that the target space is the whole real line if the IVs being considered are irrelevant, a pseudo true value if the IVs considered are endogenous but not irrelevant, and the true causal effect β in case the IVs are relevant and exogenous. This is summarized in the following proposition.

Proposition 4.1
Let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$ *then assuming that* $(y_i, x_i, z_{iS})_{i=1}^n$ *is iid such that (2.4) and (2.5) hold*

$$\begin{aligned}
&\beta_S = \mathbb{R} && \text{if} \quad \pi_S = 0 \\
&\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha && \text{if} \quad \pi_S \neq 0 \\
&\beta_S = \beta && \text{if} \quad \pi_S \neq 0 \quad \text{and} \quad \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0
\end{aligned}$$

Proof. The objective function can be decomposed in the following way

$$\begin{aligned}
\Omega(\tilde{\beta}) &\equiv \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta})) \\
&= \mathbb{E}((u_i + z'_{i\overline{E}}\alpha + x_i(\beta - \tilde{\beta}))z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(u_i + z'_{i\overline{E}}\alpha + x_i(\beta - \tilde{\beta}))) \\
&= \alpha'\mathbb{E}(z_{i\overline{E}}z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha + 2(\beta - \tilde{\beta})\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha + (\beta - \tilde{\beta})^2\pi'_S\Sigma_S\pi_S
\end{aligned}$$

Then case by case

- If $\pi_S = 0$ then

$$\Omega(\tilde{\beta}) = \alpha'\mathbb{E}(z_{i\overline{E}}z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha \Rightarrow\ \beta_S = \underset{\tilde{\beta}}{Argmin}\ \Omega(\tilde{\beta}) = \mathbb{R}$$

- If $\pi_S \neq 0$ then taking the FOC yiels

$$\beta_S : -2\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha - 2(\beta - \beta_S)\pi'_S\Sigma_S\pi_S = 0 \Leftrightarrow \beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha$$

- If $\pi_S \neq 0$ and $\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0$ then using the result in case $\pi_S \neq 0$

$$\beta_S = \beta$$

□

On the other hand if $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2)$ then the parameter target space turns out to be exactly the same, see the following proposition.

Proposition 4.2
Let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - z_{iS}'\pi_S\tilde{\beta})^2)$ *then assuming that* $(y_i, x_i, z_{iS})_{i=1}^n$ *is iid such that (2.4) and (2.5) hold*

$$\begin{aligned}
&\beta_S = \mathbb{R} && \text{if} \quad \pi_S = 0 \\
&\beta_S = \beta + (\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha && \text{if} \quad \pi_S \neq 0 \\
&\beta_S = \beta && \text{if} \quad \pi_S \neq 0 \quad \text{and} \quad \mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha = 0
\end{aligned}$$

Proof. The objective function can be decomposed in the following way

$$\begin{aligned}
\Omega(\tilde{\beta}) &\equiv \mathbb{E}((y_i - z_{iS}'\pi_S\tilde{\beta})^2) \\
&= \mathbb{E}((u_i + z_{i\overline{E}}'\alpha + v_i\beta + z_{iS}'\pi_S(\beta - \tilde{\beta}))^2) \\
&= \mathbb{E}((u_i + v_i\beta)^2) + 2(\beta - \tilde{\beta})\pi_S'\mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha + (\beta - \tilde{\beta})^2\pi_S'\Sigma_S\pi_S
\end{aligned}$$

Note that except for the first component which doesn't depend on $\tilde{\beta}$ the criterion has exactly the same decomposition as in the criterion considered in Proposition 4.1 and therefore the same solutions. □

D.2 True models in the linear IV context

In this subsection, I prove that the conditions given in section 3.1 on the level of strength a_S and the level of endogeneity b_S of the IVs in the sets $\mathcal{S}_{id}$, $\mathcal{S}_{cv}$, $\mathcal{S}_{an}$, and $\mathcal{S}_r$ are right. As long as there exists some set S such that it is exogenous and relevant then β is identified. If there is some S such that the IVs are not weak and their endogeneity level is sufficiently low relative to their strength level then 2SLS will converge. If there is some S such that the IVs are not weak and their endogeneity level is low then 2SLS will be asymptotically normal in the sense that a standard t-test confidence interval will have nominal coverage asymptotically. If there is some S such that endogeneity is sufficiently low then there exists a valid inference procedure for β. This is summarized in the following proposition.

Proposition 4.3
Assuming that $(y_i, x_i, z_i)_{i=1}^n$ *is iid such that (2.1) and (2.2) hold where for any* $S \in \mathcal{S}$, $\pi_S = n^{-a_S}\kappa_S$, $\mathbb{E}(z_{iS}z_{i\overline{E}}')\alpha = n^{-b_S}\delta_S$, $\kappa_S \in \mathbb{R}_*^S$ *is fixed and* $\delta_S \in \mathbb{R}_*^S$ *is fixed then*

- β *is identified if* $\mathcal{S}_{id} \neq \emptyset$
- *There exists some* S *such that plim* $\hat{\beta}_S = \beta$ *if* $\mathcal{S}_c \neq \emptyset$

- *There exists some S such that* $\frac{\hat{\beta}_S - \beta}{\sqrt{(x' P_{z_S} x)^{-1} \hat{\sigma}_u^2}} \xrightarrow{d} \mathcal{N}(0,1)$ *if* $\mathcal{S}_{an} \neq \emptyset$
- *There exists some S such that a valid inference method exists if* $\mathcal{S}_r \neq \emptyset$

Proof. Case by case:

- Identification directly follows from the fact that for some S, $\pi_S \neq 0$ and $\alpha_E = 0$. Indeed, for any such S, β can be expressed as $\beta = \mathbb{E}(\omega' z_{iS} x_i)^{-1} \mathbb{E}(\omega' z_{iS} y_i)$ for some non-random vector ω such that $\mathbb{E}(\omega' z_{iS} x_i) \neq 0$. Therefore, for β to be identified $\mathcal{S}_{id}$ must be non-empty.
- Recall that $\hat{C}_S \equiv \hat{\beta}_S - \beta$ then for any S such that $a_S \geqslant 1/2$

$$dlim\ \hat{C}_S \neq 0$$

which is proven in Lemma 3.1. This also prevent proper inference. Therefore, β can be consistently estimated only if there is some S such that $a_S < 1/2$.

Next, recall that $u_{iS} = z'_{i\overline{E}} \alpha + u_i$ where $\mathbb{E}(u_i | z_i) = 0$ and $\mathbb{E}(z_{iS} z'_{i\overline{E}}) \alpha = n^{-b_S} \delta_S$. Then if $a_S < 1/2$ with a slight abuse of the O_P notations

$$\begin{aligned}
\hat{C}_S = \frac{x' P_{z_S} u_S}{x' P_{z_S} x} &= n^{2a_S - 1} \frac{n^{-a_S} \kappa'_S z'_S u + n^{-a_S} \kappa_S z'_S z_{\overline{E}} \alpha + v' P_{z_S} u + v' P_{z_S} z_{\overline{E}} \alpha}{n^{-1} \kappa'_S z'_S z_S \kappa_S + 2 n^{a_S - 1/2} \kappa'_S z'_S v + n^{2a_S - 1} v P_{z_S} v} \\
&= n^{2a_S - 1} \frac{O_P(n^{1/2 - a_S}) + O_P(n^{1 - a_S - b_S}) + O_P(1) + O_P(n^{1/2 - b_S})}{O_P(1)} \\
&= \frac{O_P(n^{a_S - 1/2}) + O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1}) + O_P(n^{2a_S - b_S - 1/2})}{O_P(1)}
\end{aligned}$$

Consequently, $\hat{C}_S = o_P(1)$ if and only if $a_S < 1/2$ and $b_S - a_S > 0$ which are the conditions which characterize the sets in $\mathcal{S}_c$.

- For the t-statistic to be asymptotically normal $\hat{\beta}_S$ must be consistent, so $a_S < 1/2$, then

with a slight abuse of the O_P notations the statistic can be written as

$$\begin{aligned} t &= \frac{\hat{\beta}_S - \beta}{\sqrt{(x'P_{z_S}x)^{-1}\hat{\sigma}_u^2}} = \hat{\sigma}_u^{-1}\frac{x'P_{z_S}u_S}{\sqrt{x'P_{z_S}x}} \\ &= \hat{\sigma}_u^{-1}n^{a_S-1/2}\frac{n^{-a_S}\kappa_S'z_S'u + n^{-a_S}\kappa_S z_S'z_{\overline{E}}\alpha + v'P_{z_S}u + n^{-b_S}v'P_{z_S}z_{\overline{E}}\alpha}{\sqrt{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{a_S-1/2}\kappa_S'z_S'v + n^{2a_S-1}vP_{z_S}v}} \\ &= \hat{\sigma}_u^{-1}n^{a_S-1/2}\frac{O_P(n^{1/2-a_S}) + O_P(n^{1-a_S-b_S}) + O_P(1) + O_P(n^{1/2-b_S})}{O_P(1)} \\ &= \hat{\sigma}_u^{-1}\frac{O_P(1) + O_P(n^{1/2-b_S}) + O_P(n^{a_S-1/2}) + O_P(n^{a_S-b_S})}{O_P(1)} \end{aligned}$$

Clearly $dlim\ t = \ dlim\ \hat{\sigma}_u^{-1}\frac{\frac{1}{\sqrt{n}}\kappa_S'z_S'u}{\sqrt{\frac{1}{n}\kappa_S'z_S'z_S\kappa_S}}$ if and only if $b_S > 1/2$ and $a_S < 1/2$. Finally

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{1}{n}\sum_{i=1}^n(y_i - x_i\hat{\beta}_S)^2 = \frac{1}{n}\sum_{i=1}^n u_i^2 - \frac{2\hat{C}_S}{n}\sum_{i=1}^n u_i x_i + \frac{\hat{C}_S^2}{n}\sum_{i=1}^n x_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n u_i^2 + o_P(1) \end{aligned}$$

because $a_S < 1/2$ and $b_S - a_S > 0$ imply $\hat{C}_S = o_P(1)$. Thus, if $a_S < 1/2 < b_S$ then

$$dlim\ t = \ dlim\ \sigma_u^{-1}\frac{\frac{1}{\sqrt{n}}\kappa_S'z_S'\varepsilon_S}{\sqrt{\frac{1}{n}\kappa_S'z_S'z_S\kappa_S}} = \mathcal{N}(0,1)$$

- All weak-identification robust inference procedures are based on the fact that under $H_0 : \beta = \beta_0$ the statistic

$$S = \frac{(y - x\beta_0)'z_S(z_S'z_S)^{-1/2}}{\sqrt{\frac{1}{n}(y - x\beta_0)'M_{z_S}(y - x\beta_0)}}$$

converges in distribution towards $\mathcal{N}(0, I_s)$. Consider the numerator, with an abuse of O_P notations it can be written as

$$\begin{aligned} (y - x\beta_0)'z_S(z_S'z_S)^{-1/2} &= \frac{1}{\sqrt{n}}u_S'z_S(\frac{1}{n}z_S'z_S)^{-1/2} = \frac{1}{\sqrt{n}}u'z_S(\frac{1}{n}z_S'z_S)^{-1/2} + n^{-1/2}\alpha'z_{\overline{E}}'z_S(\frac{1}{n}z_S'z_S)^{1/2} \\ &= O_P(1) + O_P(n^{1/2-b_S}) \end{aligned}$$

Therefore, if and only if $b_S > 1/2$ can the nominator converges to a Gaussian asymptotically. With similar arguments it can be proven that the denominator converges to σ_u^2 if and only if $b_S > 0$.

Consequently, weak-identification robust inference can only be performed if there exists some S such that $b_S > 1/2$, ie if and only if $\mathcal{S}_r \neq \emptyset$.

□

D.3 Risks decomposition

In this subsection, I show prove the statements on the decomposition of the risks from section 4.1. Assumption A is maintained throughout the subsection.

The three risk can be decomposed into quadratic forms which depend on $z_{\overline{E}}^{*'}\alpha$ and $\hat{\beta}_S - \beta$. This is shown using the orthogonality between the errors (u_i, v_i) and the IVs z_i and using the independence between (y^*, x^*, z^*) and $(y_i, x_i, z_i)_{i=1}^n$, see the subset model (2.4) and (2.5).

Starting with R_{EXO}

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(y^* - x^*\hat{\beta}_S)\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((u_S^* - x^*(\hat{\beta}_S - \beta))z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(u_S^* - x^*(\hat{\beta}_S - \beta))\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((u^* + z_{\overline{E}}^{*'}\alpha - (z_S^{*'}\pi_S + v^*)(\hat{\beta}_S - \beta))z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(u^* + z_{\overline{E}}^{*'}\alpha - (z_S^{*'}\pi_S + v^*)(\hat{\beta}_S - \beta))\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta))z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta))\right)\right) \\
&= \alpha'\mathbb{E}(z_{\overline{E}}^* z_S^{*'})\Sigma_S^{-1}\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha + \mathbb{E}((\hat{\beta}_S - \beta)^2)\pi_S'\Sigma_S\pi_S - 2\pi_S'\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha \\
&= \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

Then with R_{PMSE}

$$\begin{aligned}
R_{PMSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left(y^* - z_S^{*'}\pi_S\hat{\beta}_S\right)^2\right) = \mathbb{E}\left(\mathbb{E}_n\left(x^*\beta + u_S^* - z_S^{*'}\pi_S\hat{\beta}_S\right)^2\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left(u^* + v^*\beta + z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)\right)^2\right) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

And finally with R_{MSE}

$$\begin{aligned}
R_{MSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left(y^* - x^*\hat{\beta}_S\right)^2\right) = \mathbb{E}\left(\mathbb{E}_n\left(x^*\beta + u_S^* - x^*\hat{\beta}_S\right)^2\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left(u^* - v^*(\hat{\beta}_S - \beta) + z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)\right)^2\right) \\
&= \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

Strong and endogenous IVs In case IVs subset S is strong and endogenous, as in $a_S = b_S = 0$, the difference between the IV subset estimator and β is

$$\hat{\beta}_S - \beta = \frac{\pi_S' z_S' z_{\overline{E}} \alpha}{\pi_S' z_S' z_S \pi_S} + o_P(1) = \frac{\pi_S' \mathbb{E}(z_{iS} z_{i\overline{E}}') \alpha}{\pi_S' \Sigma_S \pi_S} + o_P(1) = \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'}) \alpha}{\pi_S' \Sigma_S \pi_S} + o_P(1)$$

The proof is omitted, a more general result is derived for the proof of the main asymptotic results in appendix C. Then the risks can be rewritten as a quadratic functions of α and $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$.

For any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ the risk R_{EXO} can be rewritten as

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(||\Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2} \pi_S \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha}{\pi_S' \Sigma_S \pi_S} ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left(|| \left(\Sigma_S^{-1/2} - \frac{\Sigma_S^{1/2} \pi_S \pi_S'}{\pi_S' \mathbb{E}(z_{iS} z_{iS}') \pi_S} \right) \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left(|| \left(I_s - \frac{\Sigma_S^{1/2} \pi_S \pi_S' \Sigma_S^{1/2}}{\pi_S' \Sigma_S \pi_S} \right) \Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
&\equiv \mathbb{E}\left(||M_{\Sigma_S^{1/2}\pi_S} \Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
R_{EXO}(S) &\equiv \alpha' \mathbb{E}(z_{\overline{E}}^* z_S^{*'}) M_1 \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha + o_P(1)
\end{aligned}$$

where $M_{\Sigma_S^{1/2}\pi_S} = I_s - \Sigma_S^{1/2}\pi_S(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\Sigma_S^{1/2}$ is the projection matrix on the space orthogonal to $\Sigma_S^{1/2}\pi_S$, and $M_1 = \Sigma_S^{-1/2} M_{\Sigma_S^{1/2}\pi_S} \Sigma_S^{-1/2} = \Sigma_S^{-1} - \pi_S'(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S$ is a symmetric positive semi-definite matrix of rank $s-1$ by properties of projection matrices.

Similarly, for any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ R_{PMSE} can be rewritten as

$$\begin{aligned}
R_{PMSE}(S) &= \mathbb{E}\left((u^* + v^*\beta)^2 \right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2 \right) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2 \right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha}{\pi_S' \Sigma_S \pi_S} ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2 \right) + \mathbb{E}\left(||BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S)'\alpha ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2 \right) + \alpha' \mathbb{E}\left(BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S) BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S)' \right) \alpha + o_P(1) \\
R_{PMSE}(S) &= \mathbb{E}\left((u^* + v^*\beta)^2 \right) + \alpha' M_2 \alpha + o_P(1)
\end{aligned}$$

where $BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S) = z_{\overline{E}}^* - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi_S'\Sigma_S\pi_S)^{-1} z_S^{*'}\pi_S$ is the best linear projection of $z_{\overline{E}}^*$ on the space orthogonal to $z_S^{*'}\pi_S$, and $M_2 = \Sigma_E - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\mathbb{E}(z_S^* z_{\overline{E}}^{*'})$ is a symmetric positive semi-definite matrix by properties of projection matrices.

And finally, for any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ R_{MSE} can be rewritten as

$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \alpha' M_2 \alpha + o_P(1)$$

Exogenous IVs In case the IVs subset S is exogenous, ie if $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha = \alpha = 0$, the three risks can be rewritten as

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{PMSE}(S) = \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

Chapter 4: Nonparametric Specification Testing with SpeTestNP

Co-authored with Pascal Lavergne

Introduction

In applied work in order to evaluate the effect of a set of exogenous variables on an outcome it is very common to estimate a parametric model such as the linear model with ordinary least squares (OLS). But such parametric specifications may not capture the true relationship between outcome and exogenous variables. In fact if the chosen parametric model is a bad approximation of the true model then counterfactual analysis will be flawed. For this reason in the past forty years a literature on specification tests has developed in order to know if a parametric specification is right or wrong. `SpeTestNP` is a package which implements heteroskedasticity-robust specification tests of parametric models from Bierens (1982), Zheng (1996), Escanciano (2006), Lavergne and Patilea (2008), and Lavergne and Patilea (2012).

Hippolyte Boucher (Hippolyte.Boucher@outlook.com) is the author of `SpeTestNP` and Pascal Lavergne (lavergnetse@gmail.com) is a contributor. Both Hippolyte Boucher and Pascal Lavergne are maintainers and any question or bug should be reported to one of them. This vignette describes the principle behind each test available in `SpeTestNP`, then how to use `SpeTestNP` to test a parametric specification in practice with an illustration using the expected earnings conditional on education and age.

Testing for a parametric specification

In order to present the specification tests available in `SpeTestNP` we first describe the model being considered and define the null and alternative hypothesis, second we highlight the

principle behind each test, third we derive the test statistics and their rejection rules (based on either the bootstrap or Gaussian asymptotics), and fourth we briefly discuss and compare the tests size and power performances.

Model

Consider a sample $(y_j, x_j')_{j=1}^n$ of independent observations with y_j the scalar outcome and x_j a $k \times 1$ vector of exogenous explanatory variables. Then as long as $\mathbb{E}(|y_j|) < +\infty$ there exists some Borel-measurable regression function $g(\cdot)$ such that $g(x_j) = \mathbb{E}(y_j|x_j)$ $a.s$. That is the true model linking y_j and x_j writes

$$y_j = g(x_j) + \varepsilon_j, \qquad \mathbb{E}(\varepsilon_j|x_j) = 0 \quad a.s$$

for $j = 1, 2, \ldots, n$ and where ε_j denotes the part of y_j which is unexplained by x_j in terms of the mean. But instead in practice some parametric model characterized by a parametric family of functions $\mathcal{F} = \{f(\cdot, \tilde{\theta}) : \tilde{\theta} \in \Theta \subset \mathbb{R}^k\}$ is considered

$$y_j = f(x_j, \theta) + u_j$$

where $\theta = \underset{\tilde{\theta} \in \Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2)$ is the parameter which yields the best mean square error fit for this parametric model, and where u_j is the error induced by this parametric model. A typical estimator of θ is the non-linear least squares (NLS) estimator denoted by $\hat{\theta}$, thus when $\mathcal{F}$ is the family of linear functions then $\hat{\theta}$ is the OLS estimator. Next notice that if $g(\cdot) \in \mathcal{F}$ then $\mathbb{E}(u_j|x_j) = 0$ $a.s$ or equivalently $\mathbb{E}(y_j|x_j) = f(x_j, \theta)$. Indeed if $g(\cdot) \in \mathcal{F}$ then by properties of projections

$$g(\cdot) = \underset{\tilde{g}}{Argmin}\ \mathbb{E}((y_j - \tilde{g}(x_j))^2) = \underset{\tilde{g} \in \mathcal{F}}{Argmin}\ \mathbb{E}((y_j - \tilde{g}(x_j))^2) = \underset{\tilde{\theta} \in \Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2) = f(\cdot, \theta)$$

Consequently when modeling the true relationship between y and x with a parametric model, the implicit null hypothesis is

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s$$

And the alternative hypothesis is

$$H_1 : \mathbb{P}(\mathbb{E}(u_j|x_j) = 0) < 1$$

Equivalently the null and alternative hypothesis write

$$H_0 : g(x_j) = f(x_j, \theta) \quad a.s, \qquad H_1 : \mathbb{P}(g(x_j) = f(x_j, \theta)) < 1$$

Tests principle

Next to construct specification tests the null hypothesis is reformulated into moments conditions from which statistics can be derived. The five reformulations of the null hypothesis are in order.

Bierens (1982)

Bierens (1982) proves that the conditional moment condition of the null hypothesis is equivalent to an infinite number of moment conditions which is equivalent to an integrated conditional moment condition

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \mathbb{E}(u_j exp(i\beta' x_j)) = 0 \;\; \forall \beta \in \mathbb{R}^k \Leftrightarrow \int_{\mathbb{R}^k} \left|\mathbb{E}(u_j exp(i\beta' x_j))\right|^2 d\mu(\beta) = 0$$

where $\mu(\cdot)$ is any positive almost everywhere measure, $|\cdot|$ denotes the modulus, and i is the imaginary unit.

Zheng (1996)

Instead Zheng (1996) finds an equivalence between the conditional moment condition and an unconditional one

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \mathbb{E}(u_j \mathbb{E}(u_j|x_j) f(x_j)) = 0$$

where $f(\cdot)$ denotes the probability density function of x_j.

Escanciano (2006)

Escanciano (2006) proves the equivalence between the null hypothesis, an infinite number of moment conditions which differ from Bierens (1982), and an integrated moment condition

$$\begin{aligned} H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;&\Leftrightarrow \mathbb{E}(u_j 1\{\beta' x_j \leqslant l\}) = 0 \;\; \forall (t,l) \in \mathbb{S}^k \times \mathbb{R} \\ &\Leftrightarrow \int_{\mathbb{S}^k \times \mathbb{R}} \mathbb{E}^2(u_j 1\{\beta' x_j \leqslant l\}) f_\beta(l) d\beta dl = 0 \end{aligned}$$

where $1\{\cdot\}$ denotes the indicator function, $\mathbb{S}^k = \{\beta \in \mathbb{R}^k : |\beta| = 1\}$ denotes the unit sphere, and $f_\beta(\cdot)$ denotes the probability density function of $\beta' x_j$.

Lavergne and Patilea (2008)

Lavergne and Patilea (2008) show that the null hypothesis is equivalent to an infinite number of unconditional moment conditions

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \max_{||\beta||=1} \mathbb{E}(u_j \mathbb{E}(u_j|\beta' x_j)\omega(\beta' x_j)) = 0$$

for any $\omega(\cdot)$ such that $\forall \beta \in \mathbb{R}^k$, $\omega(\beta' x_j) > 0$ on the support of $\mathbb{E}(u_j|\beta' x_j)$. This condition resembles that of Zheng (1996) with $\beta' x_j$ replacing x_j in an effort to remove the curse of dimensionality.

Lavergne and Patilea (2012)

Finally Lavergne and Patilea (2012) prove the equivalence between the null and an integrated moment condition

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \int_B \mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j)) d\beta = 0$$

where $B \subseteq \mathbb{S}^k$ and $f_\beta(\cdot)$ denotes the density of $\beta' x_j$. This moment condition combines the integrated moments approaches of Bierens (1982) and Escanciano (2006) and the dimension reduction devise used in Lavergne and Patilea (2008).

Test statistics

Each test relies on reformulating the null hypothesis into a moment condition for which an empirical counterpart exist. Thus the test statistics are sample analogs of the moments defining the null hypothesis, possibly multiplied by the sample size in order to obtain variation at the limit. Denote by $\hat{\theta}$ a consistent estimator of θ and let $\hat{u}_j = y_j - f(x_j, \hat{\theta})$ denote the residual for individual j. The five test statistics are derived in order.

Bierens (1982)

An empirical counterpart of the integrated conditional moment $\int_{\mathbb{R}^k} \left|\mathbb{E}(u_j exp(i\beta' x_j))\right|^2 d\mu(\beta)$ of Bierens (1982) is

$$T_{icm} = \int_{\mathbb{R}^k} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{u}_j exp(i\beta' x_j) \right|^2 d\mu(\beta)$$

with some positive almost everywhere measure $\mu(\cdot)$ and where $|\cdot|$ denotes the modulus. Using properties of the modulus and of the Fourier transform it can then be shown that

$$T_{icm} = \frac{1}{n} \sum_{j,j'} \hat{u}_j \hat{u}_{j'} K(x_j - x_{j'}) = \frac{1}{n} \hat{u}' W_{icm} \hat{u}$$

where $K(\cdot)$ is the Fourier transform of $\mu(\cdot)$, $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_n)'$ is the $n \times 1$ vector of stacked residuals, and W_{icm} is the matrix with entries $K(x_j - x_{j'})$ for any row j and column j'. Although this statistic can be used as is, $\mu(\cdot)$ is typically assumed to be a symmetric probability measure which is strictly positive almost everywhere. This simplifies the asymptotic theory and the derivation of the test statistic in practice. Indeed as a consequence the Fourier transform of $\mu(\cdot)$ denoted as $K(\cdot)$ is a symmetric bounded density. Hence candidates for $K(\cdot)$ include logistic, triangular, normal, student, or Cauchy densities, see Johnson, Kotz and Balakrishnan (1995, section 23.3) and Dreier and Kotz (2002). Furthermore to control for scale, we impose that either the integral of $K(\cdot)$ to the square equals one or that the distribution associated to $K(\cdot)$ has variance one.

Zheng (1996)

Zheng (1996) test statistic is the sample analog of $\mathbb{E}(u_j \mathbb{E}(u_j|x_j)f(x_j))$ which is derived by estimating both the density $f(\cdot)$ of x_j and the conditional mean $\mathbb{E}(u_j|x_j = \cdot)$ with Kernels. For any $\tilde{x} \in \mathbb{R}^k$ define

$$\hat{f}(\tilde{x}) = \frac{1}{nh^k} \sum_j K\left(\frac{\tilde{x} - x_j}{h}\right), \qquad \hat{\mathbb{E}}(u_j|x_j = \tilde{x}) = \frac{1}{nh^k} \sum_j \frac{u_j}{\hat{f}(\tilde{x})} K\left(\frac{\tilde{x} - x_j}{h}\right)$$

where $K(\cdot)$ is a Kernel function which is nonnegative, symmetric, bounded, continuous and which integrates to one, and h a bandwidth such that $h \underset{n\to+\infty}{\to} 0$ and $nh^k \underset{n\to+\infty}{\to} +\infty$. Then the test statistic is the sample analog of the moment $\mathbb{E}(u_j \mathbb{E}(u_j|x_j)f(x_j))$

$$T_{zheng} = \frac{1}{n} \sum_j \hat{u}_j \hat{\mathbb{E}}(u_{j'}|x_{j'} = x_j)\hat{f}(x_j)$$

It can be rewritten as

$$T_{zheng} = \frac{1}{n(n1)h^k} \sum_{j,j\neq j} \hat{u}_j \hat{u}_j K\left(\frac{x_j x_j}{h}\right) = \frac{1}{n(n1)h^k} \hat{u}^{W_{zheng}\hat{u}}$$

where W_{zheng} is a matrix whose diagonal elements are equal to zero and its other entries are equal to $K\left(\frac{x_j x_j}{h}\right)$ for any row j any column j such that $j \neq j$.

Escanciano (2006)

Escanciano (2006) test statistic is the sample analog of $\int_{\mathbb{S}^k \times \mathbb{R}} \mathbb{E}^2(u_j 1\{\beta' x_j \leqslant l\}) f_\beta(l) d\beta dl$ times n which is derived by approximating the density $f_\beta(\cdot)$ by a probability mass function. Let $\hat{f}_\beta(l) = \frac{1}{n} \sum_r 1\{\beta' x_r = l\}$ then the statistic is

$$T_{esca} = \int_{\mathbb{S}^k \times \mathbb{R}} \left(\frac{1}{\sqrt{n}} \sum_j \hat{u}_j 1\{\beta' x_j \leqslant l\}\right)^2 \hat{f}_\beta(l) d\beta dl$$

It can be proven that it has the same form as the other test statistics

$$T_{esca} = \frac{1}{n}\sum_{j,j'}\hat{u}_j\hat{u}_{j'}\frac{1}{n}\sum_r\int_{\mathbb{S}^k}1\{\beta'x_j \leqslant \beta'x_r, \beta'x_{j'} \leqslant \beta'x_r\}d\beta = \frac{1}{n}\hat{u}'W_{esca}\hat{u}$$

where W_{esca} has elements $\frac{1}{n}\sum_r W_{esca,j,j',r}$ with $W_{esca,j,j',r} = \int_{\mathbb{S}^k}1\{\beta'x_j \leqslant \beta'x_r, \beta'x_{j'} \leqslant \beta'x_r\}d\beta$ for any row j and column j'. Approximating the integrals in W_{esca} is unnecessary because

$$W_{esca,j,j',r} = W^{(0)}_{esca,j,j',r}\frac{\pi^{k/2}-1}{\Gamma(k/2+1)}, \qquad W^{(0)}_{esca,j,j',r} = \left|\pi - arccos\left(\frac{(x_j-x_r)'(x_{j'}-x_r)}{|x_j-x_r||x_{j'}-x_r|}\right)\right|$$

See appendix B in Escanciano (2006) for more details. Note that n^3 operations are necessary to compute W_{esca} which means that this statistic takes much more time to compute.

Lavergne and Patilea (2008)

Lavergne and Patilea (2008) consider a sample analog of the moment $\mathbb{E}(u_j\mathbb{E}(u_j|x_j)\omega(\beta'x_j))$ and replace $\omega(\cdot)$ by $f_\beta(\cdot)$ the density of $\beta'x_j$. In addition they replace β by the value in the unit hypersphere which maximizes the moment taken to the square. This way the test is given the direction which best reject the null hypothesis under the alternative. Thus first define for any $t \in \mathbb{S}^k$

$$\mathcal{Q}(\beta) = \frac{1}{n(n-1)h}\sum_{j,j'\neq j}\hat{u}_j\hat{u}_{j'}K\left(\frac{\beta'(x_j-x_{j'})}{h}\right)$$

where $K(\cdot)$ is a bounded symmetric density with bounded variation, h is a bandwidth such that $h \underset{n\to+\infty}{\longrightarrow} 0$ and $\frac{(nh^2)^\delta}{log(n)} \underset{n\to+\infty}{\longrightarrow} +\infty$ for some $\delta \in (0;1)$. $\mathcal{Q}(\beta)$ cannot be directly used, instead define $\hat{\beta}$ the direction which best captures the correlation between the residuals and the explanatory variables

$$\hat{\beta} = \underset{\beta\in\mathbb{S}^k}{Argmax}\ |n\sqrt{h}\mathcal{Q}(\beta)\alpha_n 1\{\beta \neq \beta^*\}|$$

where β^* represents a favored direction chosen a priori, and $\alpha_n \underset{n\to+\infty}{\to} 0$ is the weight given to this favored direction. β^* and α_n improve significantly the power properties of the test in small

sample. Note that in practice the unit hypersphere $\mathbb{S}^k$ is approximated by a finite number of points. Thus the test statistic is the criterion evaluated at $\hat{\beta}$

$$T_{pala} = \mathcal{Q}(\hat{\beta}) = \frac{1}{n(n1)h} \sum_{j,j \neq j} \hat{u}_j \hat{u}_j K\left(\frac{\hat{\beta}(x_j x_j)}{h}\right) = \frac{1}{n(n1)h} \hat{u}^{W_{pala}\hat{u}}$$

where W_{pala} is a matrix with diagonal elements equal to zero and its other entries equal to $K\left(\frac{\hat{(x_j x_j)}}{h}\right)$ for any row j and column j such that $j \neq j$.

Lavergne and Patilea (2012)

Finally, Lavergne and Patilea (2012) use the sample analog of $\int_B \mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j))d\beta = 0$ for some $B \subseteq \mathbb{S}^k$ as a test statistic. To derive it notice that an empirical counterpart of $\mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j))$ is $\mathcal{Q}(\beta)$ as defined in previously. Hence, their test statistic which they call smooth integrated conditional moment statistic writes

$$T_{sicm} = \int_B \mathcal{Q}(\beta)d\beta = \int_B \frac{1}{n(n-1)h} \sum_{j,j' \neq j} \hat{u}_j \hat{u}_{j'} K\left(\frac{\beta'(x_j - x_{j'})}{h}\right) d\beta = \frac{1}{n(n-1)h} \hat{u}' W_{sicm} \hat{u}$$

where W_{sicm} has diagonal elements equal to zero and its other elements are equal to $\int_B K\left(\frac{\beta'(x_j - x_{j'})}{h}\right) d\beta$ for any row j and any column $j' \neq j$. Clearly T_{sicm} is a smooth version of T_{icm} because of the bandwidth h. Furthermore, it is also a smooth version of T_{pala} in the sense that instead of being based on the squared error in the worst direction of $\beta' x_j$, it is based on a continuum of directions. In practice, to compute the integral a finite number of points are drawn randomly from B and B doesn't have to be the whole unit hypersphere $\mathbb{S}^k$. For instance, half hyperspheres can be considered such as $\{\beta \in \mathbb{R}^k : \beta_m \geqslant 0, ||\beta|| = 1\}$ where β_m denotes the m-th element of the vector β.

Normalization

The five test statistics can be normalized. Not only does this improve the finite sample properties of the tests, but it allows to use Gaussian asymptotics when deciding to reject the null hypothesis with the tests of Zheng (1996), Lavergne and Patilea (2008), and Lavergne and Patilea (2012). This is extremely useful in large samples instead of using the bootstrap.

The normalized test statistics are of the following form:

$$\hat{T}_{icm} = \hat{u}\hat{W}_{icm}\hat{u}, \qquad \hat{W}_{icm} = W_{icm}\sqrt{2\sum_{j,j}\hat{\sigma}_j^2\hat{\sigma}_j^2 K^2(x_j x_j)}$$

$$\hat{T}_{zheng} = \hat{u}\hat{W}_{zheng}\hat{u}, \qquad \hat{W}_{zheng} = W_{zheng}\sqrt{2\sum_{j,j\neq j}\hat{\sigma}_j^2\hat{\sigma}_{j'}^2 K^2\left(\frac{x_j x_j}{h}\right)}$$

$$\hat{T}_{esca} = \hat{u}\hat{W}_{esca}\hat{u}, \qquad \hat{W}_{esca} = W_{esca}\sqrt{2\sum_{j,j}\hat{\sigma}_j^2\hat{\sigma}_j^2\left(\frac{1}{n}\sum_r\int_{S^k} 1\{\beta x_j \leqslant \beta x_r, \beta x_j \beta x_r\}d\beta\right)^2}$$

$$\hat{T}_{pala} = \hat{u}\hat{W}_{pala}\hat{u}, \qquad \hat{W}_{pala} = W_{pala}\sqrt{2\sum_{j,j\neq j}\hat{\sigma}_j^2\hat{\sigma}_{j'}^2 K^2\left(\frac{\gamma(x_j x_j)}{h}\right)}$$

$$\hat{T}_{sicm} = \hat{u}\hat{W}_{sicm}\hat{u}, \qquad \hat{W}_{sicm} = W_{sicm}\sqrt{2\sum_{j,j\neq j}\hat{\sigma}_j^2\hat{\sigma}_{j'}^2\left(\int_B K\left(\frac{\beta(x_j x_j)}{h}\right)d\beta\right)^2}$$

where $\hat{\sigma}_j^2$ controls for the conditional variance of the error uj. A naive approach to the normalization which works very well in large sample is to directly replace $\hat{\sigma}_j^2$ by the squared residuals $\hat{u}_j^2$. Another approach to the normalization is to replace $\hat{\sigma}_j^2$ by an estimator such the as the nonparametric kernel variance estimator of Yin, Geng, Li and Wang (2010) which writes

$$\hat{\sigma}^2(\tilde{x}) = \frac{\frac{1}{nh_v}\sum_j (y_j\overline{y}(\tilde{x}))^2 K\left(\frac{\tilde{x}x_j}{h_v}\right)}{\frac{1}{nh_v}\sum_j K\left(\frac{\tilde{x}x_j}{h_v}\right)}, \qquad \overline{y}(\tilde{x}) = \frac{\frac{1}{nh_v}\sum_j y_j K\left(\frac{\tilde{x}x_j}{h_v}\right)}{\frac{1}{nh_v}\sum_j K\left(\frac{\tilde{x}x_j}{h_v}\right)}$$

where K is a Kernel function and h_v is a bandwidth which can be different from h.

Both the naive and nonparametric approaches to the normalization are implemented.

Rejection rules

To decide whether to reject or not the null hypothesis we need to compute quantiles of the distribution of each statistic under the null conditional on $x = (x_1, \ldots, x_n)$. Then H_0 is rejected

at level 5% if the test statistic is above the quantile 95% of its distribution under the null. To compute these quantiles we propose two solutions.

First we consider computing the quantiles using the fixed design bootstrap. x is held fixed so for each test statistic their central W is held fixed, and a n×1 vector of residuals ^ub is drawn using the fixed design wild bootstrap of Wu (1986) or the smooth conditional moment bootstrap of Gozalo (1997). It will also control for potential heteroskedasticity. Using this bootstrapped vector of residuals and the maintained central matrix W a bootstrapped statistic can be computed. After repeating this operation many times we obtain a vector of bootstrapped statistics. The quantiles of this vector can then be used to reject or not H_0. As an example if the test we consider is that of Bierens (1982) a bootstrapped statistic is

$$T_{icm,b} = \frac{1}{n}\hat{u}_b W_{icm}\hat{u}_b$$

By repeating this operation B times we obtain B bootstrapped statistics $(T_{icm,b})_{b=1}^{B}$ which mimic the behavior of T_{icm} under the null hypothesis. Consequently the parametric specification will be rejected at level 5% if $T_{icm} > q_{95\%}$ where $q_{95\%}$ is the 95% quantile of $(T_{icm,b})_{b=1}^{B}$. The same procedure can be applied to other tests and their normalized versions to decide whether or not to reject the null hypothesis.

Second we consider using the quantiles of the standard normal. As mentioned, the normalized versions of the statistics of Zheng (1996), Lavergne and Patilea (2008), and Lavergne and Patilea (2012) are asymptotically standard normal. Thus if one of these normalized test statistics are used, we can use the quantiles of a standard normal to reject or not H_0. As an example if the test we consider is that of Zheng (1996) with a normalization then the parametric specification will be rejected at level 5% if $|\hat{T}_{zheng}| > 1.96$.

Validity, consistency and power properties

Each test can be proven to be valid, as in under the null hypothesis the probability to reject the null converges to nominal level, and to be consistent, as in under any fixed alternative the probability to reject the null converges to one.

But these five tests differ significantly in terms of power in practice. The test of Zheng (1996) seem to be the least powerful test in practice, it has no power against Pitman alternatives and has difficulty rejecting the null when the number k of exogenous variables is large. The test of

Bierens (1982) possesses more than trivial power against Pitman alternatives but it also has trouble rejecting the null when k is large. The test of Escanciano (2006) does not depend on a choice of weighting function and does not require numerical integration however to derive its statistic it requires n^3 operations making it very slow and hard to apply in practice. In addition its power however largely depends on the true alternative and is low when k is large. The tests of Lavergne and Patilea (2008), and Lavergne and Patilea (2012) are more powerful than the other two when k is large because of their use of a continuum of single index β^{x_j} to summarize the correlation between u_j and x_j. At the same time when k is small the two tests are at least as powerful as the others. As mentioned the power of Lavergne and Patilea (2008) test comes from the "worst" single-index alternative whereas the power of Lavergne and Patilea (2012) test comes from a continuum of single-index alternatives. Thus in practice under the alternative the nature of the correlation between u_j and x_j will determine which of these two tests is more powerful.

See the references for more details.

Using `SpeTestNP`

Previously we have described the principle behind the five nonparametric specification tests, how to derive the test statistics and the rejection rules, and discussed their properties. Next we show how to use `SpeTestNP` to test parametric models in practice, with first the installation, second a description of how to use the test, third a thorough description of the arguments of the package main function `SpeTest`, and fourth an illustration to determine the true shape of expected wages conditional on years of education and age.

Installation

To install `SpeTestNP` from CRAN simply run the following command:

```
install.packages("SpeTestNP")
```

To install `SpeTestNP` from Github the package `devtools` should be installed and the following commands should be run:

```
install.packages("devtools")
```

```
library("devtools")


install_github("HippolyteBoucher/SpeTestNP")
```

To choose where and how the package is installed check `help(install_github)` and `help(install.packages)`. Alternatively users can download the package and directly install it with the CMD. `SpeTestNP` requires the packages `stats` (already installed and loaded by default in Rstudio), `foreach`, `parallel` and `doParallel` (if parallel computing is used to generate the vector) to be installed.

Testing with `SpeTestNP`

Recall the true model and the model induced by the parametric specification characterized by $\mathcal{F} = \{f(\cdot, \tilde{\theta}) : \tilde{\theta} \in \Theta \subset \mathbb{R}^k\}$

$$y_j = g(x_j) + \varepsilon_j, \qquad y_j = f(x_j, \theta) + u_j$$

where $\mathbb{E}(y_j|x_j) = g(x_j)$ *a.s* and $\theta = \underset{\tilde{\theta}\in\Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2)$.

Then to test the parametric specification or equivalently to test $H_0 : \mathbb{E}(u_j|x_j) = 0$ *a.s* the function `SpeTest` of the package `SpeTestNP` can be directly used by filling the first argument `eq` with a fitted model of class `lm` or `nls`. In case the parametric specification is linear or can be rewritten in a linear form `eq` should be an object of class `lm`. In case of non-linear models `eq` should be an object of class `nls` which stands for non-linear least squares (from the package `stats`). Note that in order to perform the specification test by feeding `SpeTest` with an `nls` model then the arguments in `nls` must be given in the right order. Then by running the following command the parametric specification characterized by $\mathcal{F}$ is tested

```
  SpeTest(eq)
```

The function returns an object of class `STNP` which when printed with `print` or `print.STNP` returns the test statistic and its p-value. An object of type `STNP` is a list which not only contains the test statistic `stat` and its p-value `pval` but also the type of the test `type`, the rejection rule `rejection`, the test statistic normalization `norma`, the Kernel function denoted as $K(\cdot)$ used to compute the test statistic central matrix `ker`, the standardization method of test the statistic central matrix `knorm`, the type of bootstrap used to compute the p-value `boot`, the number of

bootstrap samples used to compute the p-value `nboot`, the bandwidths `cch` and `hv`, etc... To obtain a summary of the test and its options the method `summary` or `summary.STNP` can be used on objects of class `STNP`.

By default the test of Bierens (1982) with the standard normal density as the central matrix function is applied and the test p-value is obtained using 50 wild bootstrap samples with a naive estimator of the conditional variance of the errors. Among many options, by changing the argument `rejection` from `bootstrap` (the default) to `asymptotics` if `type = "zheng"` or `type = "pala"` or `type = "sicm"` the test p-value is then based on the asymptotic normality of these normalized test statistics under the null. In addition by default the test statistic is not normalized as in by default the denominator in T_{zheng}, T_{pala} and T_{sicm} is set to one. This can be changed by setting `norma = "naive"` to normalize the statistic using a naive estimator of the errors conditional variance, or by setting `norma = "np"` to normalize the statistic using a nonparametric estimator of the errors conditional variance. If `rejection = "bootstrap"` setting `para` to `TRUE` greatly speeds up the computation of the p-value by deriving bootstrapped statistics in parallel. For more details refer to the next section or `help(SpeTest)`.

Note that the functions `SpeTest_Stat` and `SpeTest_Dist` are also available. Both functions take similar arguments to `SpeTest`. `SpeTest_Stat` computes the specification test statistic, while `SpeTest_Dist` generates a vector of size `nboot` from the specification test statistic distribution under the null hypothesis using the bootstrap. The argument `para` is also available to `SpeTest_Dist`. `SpeTest_Stat` and `SpeTest_Dist` allow to easily perform simulation exercises.

Arguments description and additional features

To be more specific about the arguments of the function `SpeTest`:

- Argument `eq` should be the fitted parametric model of class `lm` or `nls`of the parametric specification of interest $\mathcal{F}$
- Argument `type` refers to the type of the test

  If `type = "icm"` the test of Bierens (1982) is performed (default)

  If `type = "zheng"` the test of Zheng (1996) is performed

  If `type = "esca"` the test of Escanciano (2006) is performed, significantly increases computing time

If `type = "pala"` the test of Lavergne and Patilea (2008) is performed

If `type = "sicm"` the test of Lavergne and Patilea (2012) is performed

- Argument `rejection` refers to the rejection rule

  If `rejection = "bootstrap"` the p-value of the test is based on the bootstrap (default)

  If `rejection = "asymptotics"` and `type = "zheng"` or `type = "esca"` or `type = "sicm"` the p-value of the test is based on asymptotic normality of the normalized version of one of these test statistic under the null hypothesis

  If `type = "icm"` or `type = "esca"` the argument `rejection` is ignored and the p-value is based on the bootstrap

- Argument `norma` refers to the normalization of the test statistic

  If `norma = "no"` the test statistic is not normalized (default)

  If `norma = "naive"` the test statistic is normalized with a naive estimator of the errors variance

  If `norma = "np"` the test statistic is normalized with a nonparametric estimator of the errors variance

- Argument `boot` refers to the bootstrap method used to compute the test p-value when `rejection = "bootstrap"`

  If `boot = "wild"` the wild bootstrap of Wu (1986) is used (default)

  If `boot = "smooth"` the smooth conditional moments bootstrap of Gozalo (1997) is used

- Argument `nboot` is the number of bootstraps used to compute the test p-value, by default ‘nboot = 50}

- Argument `para` determines if parallel computing is used or not when `rejection = "bootstrap"`

  If `para = FALSE` parallel computing is not used to generate the bootstrap samples to compute the test p-value (default)

  If `para = TRUE` parallel computing is used to generate the bootstrap samples to compute the test p-value, significantly decreases computing time, makes use of all CPU cores

except one

- Argument `ker` refers to the Kernel function used in the central matrix and for the nonparametric covariance estimator if there is any

  If `ker = "normal"` the central matrix Kernel function is the normal p.d.f (default)

  If `ker = "triangle"` the central matrix Kernel function is the triangular p.d.f

  If `ker = "logistic"` the central matrix Kernel function is the logistic p.d.f

  If `ker = "sinc"` the central matrix Kernel function is the sine cardinal function

- Argument `knorm` refers to the normalization of the Kernel function

  If `knorm = "sd"` then the standard deviation using the Kernel function equals 1 (default)

  If `knorm ="sq"` then the integral of the squared Kernel function equals 1

- Argument `cch` is the central matrix Kernel bandwidth

  If `type = "icm"` or `type = "esca"` then `cch` always equals `1`

  If `type = "zheng"` the `"default"` bandwidth is the scaled rule of thumb: `cch = 1.06*n^(-1/5)`

  If `type = "sicm"` and `type = "pala"` the `"default"` bandwidth is the scaled rule of thumb: `cch = 1.06*n^(-1/(4+k))` where `k` is the number of regressors

  The user may change the bandwidth when `type = "zheng"`, `type = "sicm"` or `type = "pala"`.

- Argument `hv` is the bandwidth the nonparametric errors covariance estimator when `norma = "np"` or `rejection = "bootstrap"` and `boot = "smooth"`

  By `"default"` the bandwidth is the scaled rule of thumb `hv = 1.06*n^(-1/(4+k))`

- Argument `nbeta` refers to the number of elements β used to represent the unit hypersphere $\mathcal{S}^k$ when `type = "pala"` or `type = "sicm"`

  Computing time increases as `nbeta` gets larger

  By `"default"` it is equal to 20 times the square root of the number of exogenous control variables

- Argument `direct` refers to the default "directions" for the tests of Lavergne and Patilea (2008) and Lavergne and Patilea (2012)

  If `type = "pala"`, `direct` is the favored direction for β, by `"default"` it is the OLS estimator if `class(eq) = "lm"`

  If `type = "sicm"`, `direct` is the initial direction for β. This direction should be a vector of `0` (for no direction), `1` (for positive direction) and `-1` (for negative direction)

  For example, `c(1,-1,0)` indicates that the user thinks that the 1st regressor has a positive effect on the dependent variable, that the 2nd regressor has a negative effect on the dependent variable, and that he has no idea about the effect of the 3rd regressor

  By `"default"` no direction is given to the hypersphere

- Argument `alphan` refers to the weight given to the favored direction for β when `type = "pala"`

  By `"default"` it is equal to `log(n)*n^(-3/2)`

Before changing the default options of arguments `norma`, `direct` and `alphan` we strongly advise the user to read the tests references.

Illustration

To finish we use data on 1,000 individuals from the Current Population Survey as in Stock and Watson (2007) to find the true shape of their expected earnings conditional on their years of education and their age using the test of Bierens (1982).

```
library(SpeTestNP)
library(AER)

### Loading the data and taking a first look

data( CPSSW8 )

summary ( CPSSW8 )
```

```
#>     earnings         gender           age              region
```

```
#>  Min.   : 2.003   male  :34348   Min.   :21.00   Northeast:12371
#>  1st Qu.:11.058   female:27047   1st Qu.:33.00   Midwest  :15136
#>  Median :16.250                  Median :41.00   South    :18963
#>  Mean   :18.435                  Mean   :41.23   West     :14925
#>  3rd Qu.:23.558                  3rd Qu.:49.00
#>  Max.   :72.115                  Max.   :64.00
#>    education
#>  Min.   : 6.00
#>  1st Qu.:12.00
#>  Median :13.00
#>  Mean   :13.64
#>  3rd Qu.:16.00
#>  Max.   :20.00
```

Thus the dependent variable we consider is earnings and the explanatory variables we use to build the conditional expectation are education and age. First we fit a linear specification of conditional earnings.

```
    lm_lin <- lm( earnings ~ age + education,
                      data = CPSSW8[1:1000,] )

    summary ( lm_lin )
```

```
#>
#> Call:
#> lm(formula = earnings ~ age + education, data = CPSSW8[1:1000,
#>     ])
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -27.313  -6.464  -1.445   4.804  42.092
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -14.18639    2.10661  -6.734 2.78e-11 ***
#> age           0.15846    0.02747   5.767 1.07e-08 ***
```

```
#> education     1.93904    0.12286  15.782  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 9.465 on 997 degrees of freedom
#> Multiple R-squared:  0.2176, Adjusted R-squared:  0.216
#> F-statistic: 138.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

Both variables are very significant. Then we perform two tests of the linear specification, the bootstrap test of Bierens (1982) using the bootstrap decision rule, and the asymptotic test of Zheng (1996) with a naive normalization.

```
    SpeTest( lm_lin , type = "icm" , rejection = "bootstrap" )
```

```
#>
#>   Bierens (1982) integrated conditional moment test
#>
#>   Test statistic :  27.31333
#>   Bootstrap p-value :  0
#>
```

```
    SpeTest( lm_lin , type = "zheng" , rejection = "asymptotics" )
```

```
#>
#>   Zheng (1996) test
#>
#>   Normalized test statistic :  1.47353
#>   Asymptotic p-value :  0.0703
#>
```

The linear specification is rejected at level below 1% for the test of Bierens (1982) and at level below 10% for the test of Zheng (1996). So we fit a quadratic specification and perform the same tests.

```
    lm_quad <- lm( earnings ~ age + I(age^2) + education + I(education^2),
                      data = CPSSW8[1:1000,] )

    summary( lm_quad )
```

```
#>
#> Call:
#> lm(formula = earnings ~ age + I(age^2) + education + I(education^2),
#>     data = CPSSW8[1:1000, ])
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -32.167  -6.242  -1.412   4.665  41.753
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    -3.353005   8.633125  -0.388  0.69781
#> age             1.011953   0.212083   4.772  2.1e-06 ***
#> I(age^2)       -0.010051   0.002456  -4.093  4.6e-05 ***
#> education      -2.079218   1.041245  -1.997  0.04611 *
#> I(education^2)  0.140968   0.036501   3.862  0.00012 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 9.323 on 995 degrees of freedom
#> Multiple R-squared:  0.2424, Adjusted R-squared:  0.2393
#> F-statistic: 79.58 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
    SpeTest( lm_quad , type = "icm" , rejection = "bootstrap" )
```

```
#>
#>   Bierens (1982) integrated conditional moment test
#>
#>   Test statistic :  1.45746
#>   Bootstrap p-value :  0.18
#>
```

```
    SpeTest( lm_quad , type = "zheng" , rejection = "asymptotics")
```

```
#>
#>   Zheng (1996) test
```

```
#>
#>   Normalized test statistic :  -0.98736
#>   Asymptotic p-value :  0.16173
#>
```

Both age and education to the square are very significant. In addition the p-values of both tests are above 15% so we cannot reject the quadratic specification. Finally we test a highly non-linear specification with age, age to the square, education, education to the square, and their products included as controls:

```
    lm_nlin <- lm( earnings ~ age + I(age^2) + education + I(education^2)
                      + I(education*age) + I(education^2*age)
                      + I(education*age^2) + I(education^2*age^2),
                      data= CPSSW8[1:1000,] )

    summary( lm_nlin )
```

```
#>
#> Call:
#> lm(formula = earnings ~ age + I(age^2) + education + I(education^2) +
#>     I(education * age) + I(education^2 * age) + I(education *
#>     age^2) + I(education^2 * age^2), data = CPSSW8[1:1000, ])
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -33.135  -6.212  -1.485   4.515  41.920
#>
#> Coefficients:
#>                           Estimate Std. Error t value Pr(>|t|)
#> (Intercept)             6.006e+01  1.334e+02   0.450    0.653
#> age                    -3.545e-01  6.060e+00  -0.058    0.953
#> I(age^2)               -1.043e-02  6.707e-02  -0.155    0.876
#> education              -9.404e+00  1.924e+01  -0.489    0.625
#> I(education^2)          3.335e-01  6.815e-01   0.489    0.625
#> I(education * age)      1.277e-01  8.738e-01   0.146    0.884
#> I(education^2 * age)   -2.053e-03  3.093e-02  -0.066    0.947
```

```
#> I(education * age^2)    6.633e-04  9.669e-03   0.069    0.945
#> I(education^2 * age^2) -4.410e-05  3.420e-04  -0.129    0.897
#>
#> Residual standard error: 9.316 on 991 degrees of freedom
#> Multiple R-squared:  0.2467, Adjusted R-squared:  0.2406
#> F-statistic: 40.56 on 8 and 991 DF,  p-value: < 2.2e-16
```

```
    SpeTest( lm_nlin , type = "icm" , rejection = "bootstrap" )
```

```
#>
#>   Bierens (1982) integrated conditional moment test
#>
#>   Test statistic :  0.02541
#>   Bootstrap p-value :  0.64
#>
```

```
    SpeTest( lm_nlin , type = "zheng" , rejection = "asymptotics")
```

```
#>
#>   Zheng (1996) test
#>
#>   Normalized test statistic :  -1.8227
#>   Asymptotic p-value :  0.03417
#>
```

This time none of the variables are considered (individually) significant. This does not mean that this specification is wrong, in fact it nests the quadratic specification. Note that the p-value of the test of Bierens (1982) is very high while the p-value of asymptotic test of Zheng (1996) is 3%. This difference can be explained by the fact that both tests have important size distortions when the number of explanatory variables is "large". Thus we perform a final check with the asymptotic tests of Lavergne and Patilea (2008) and Lavergne and Patilea (2012).

```
    SpeTest( lm_nlin, type = "pala", rejection = "asymptotics", nbeta = 40 )
```

```
#>
#>   Lavergne and Patilea (2008) test
#>
#>   Normalized test statistic :  -0.80158
```

```
#>   Asymptotic p-value :  0.2114
#>
```

```
    SpeTest( lm_nlin, type = "pala", rejection = "bootstrap" , nboot = 10 , nbeta = 10 )
```

```
#>
#>   Lavergne and Patilea (2008) test
#>
#>   Test statistic :  -128.10515
#>   Bootstrap p-value :  0.5
#>
```

Both p-values are high so we cannot reject this highly non-linear specification.

References

H.J. Bierens (1982), "Consistent Model Specification Test", *Journal of Econometrics*, 20 (1), 105-134

I. Dreier and S. Kotz (2002), "A note on the characteristic function of the t-distribution", *Statistics & Probability Letters*, 57 (3), 221-224

J.C. Escanciano (2006), "A Consistent Diagnostic Test for Regression Models Using Projections", *Econometric Theory*, 22 (6), 1030-1051

P.L. Gozalo (1997), "Nonparametric Bootstrap Analysis with Applications to Demographic Effects in Demand Functions", *Journal of Econometrics*, 81 (2), 357-393

Johnson, Kotz and Balakrishnan (1995), "Continuous Univariate Distributions", volume 2, *Wiley Series in Probability and Statistics: Applied Probability and Statistics*, Wiley & Sons

P. Lavergne and V. Patilea (2008), "Breaking the Curse of Dimensionality in Nonparametric Testing", *Journal of Econometrics*, 143 (1), 103-122

P. Lavergne and V. Patilea (2012), "One for All and All for One: Regression Checks with Many Regressors", *Journal of Business & Economic Statistics*, 30 (1), 41-52

J.H. Stock and M.W. Watson (2006), "Why Has U.S. Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking*, 39 (1), 3-33

C.F.J. Wu (1986) “Jackknife, bootstrap and other resampling methods in regression analysis (with discussion)”, *National Bureau of Economic Research Working Paper*

J. Yin, Z. Geng, R. Li, H. Wang (2010), “Nonparametric covariance model”, *Statistica Sinica*, 20 (1), 469-479

J.X. Zheng (1996), “A Consistent Test of Functional Form via Nonparametric Estimation Techniques”, *Journal of Econometrics*, 75 (2), 263-289